DOCUMENT RESUME

ED 092 920 CS 001 200

AUTHOR	Carver, Ronald P.
TITLE	Improving Reading Comprehension: Measuring Readability. Final Report.
INSTITUTION	American Institutes for Research in the Behavioral Sciences, Silver Spring, Md.
SPONS AGENCY	Office of Naval Research, Washington, D.C. Personnel and Training Research Programs Office.
REPORT NO	AIR-30801-5-74-FR
PUB DATE	May 74
NOTE	90p.
EDRS PRICE	MF-$0.75 HC-$4.20 PLUS POSTAGE
DESCRIPTORS	Elementary Education; *Measurement Techniques; *Readability; Reading Ability; *Reading Comprehension; Reading Difficulty; *Reading Improvement; Reading Materials; *Reading Research
IDENTIFIERS	*Rauding Scale

ABSTRACT

A standardized method, called programed prose, has been developed which can be used to automatically convert prose training material into a form which forces trainees to read the material with at least a minimal level of comprehension. From the results of a series of experimental studies, it was concluded that programed prose facilitates learning under the following conditions: (1) when individuals are not always highly motivated to learn; (2) when attention to the reading task wanes; (3) when the reading ability level of the individual exceeds the reading difficulty level of the prose material; and (4) when the time allowed for reading exceeds the time necessary to complete the programed prose. It was also found that programed prose is more effective and more efficient in facilitating learning than is the study question technique that is used in correspondence course material, and it was concluded that the Rauding Scale was more valid as a measure of readability than either the Flesch or the Dale-Chall measures. (Author/RB)

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

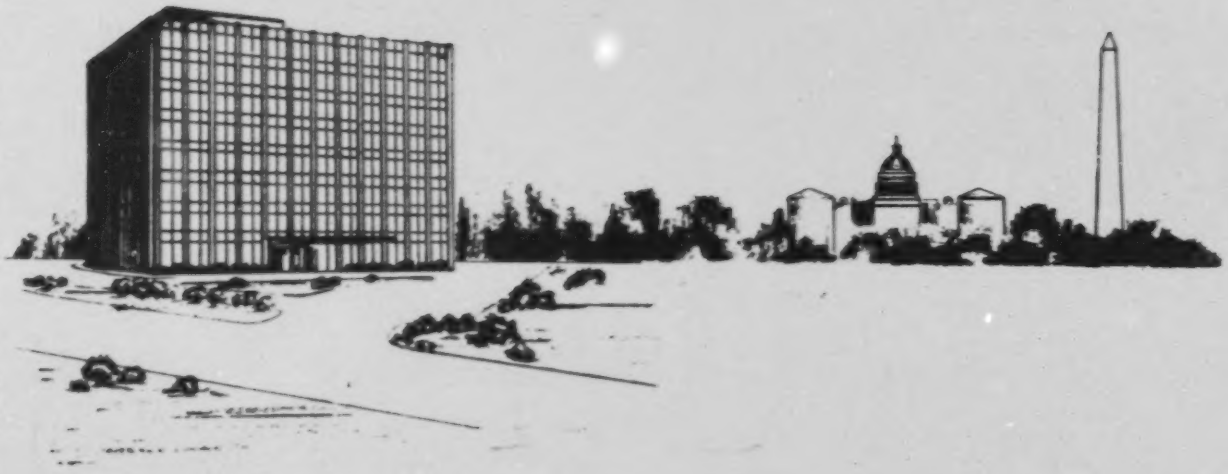
ED 092920

CS

Improving Reading Comprehension: Measuring Readability

Ronald P. Carver

Final Report
MAY 1974



CS 001 200

AMERICAN INSTITUTES FOR RESEARCH
WASHINGTON OFFICE
Address: 8555 Sixteenth Street, Silver Spring, Maryland 20910
Telephone: (301) 587-8201

R74-2

AMERICAN INSTITUTES FOR RESEARCH

WASHINGTON, D.C.

EDWIN A. FLEISHMAN, PhD, DIRECTOR

Albert S. Glickman, PhD, Deputy Director

HUMAN RESOURCES RESEARCH GROUP

Clifford P. Hahn, MS, Director

Studies on personnel selection, training, instructional and training methods, proficiency measurement, accidents, and evaluation of educational and social programs.

ORGANIZATIONAL BEHAVIOR RESEARCH GROUP

Albert S. Glickman, PhD, Director

Research on individual, interpersonal and group behavior as they relate to organizational functioning and effectiveness, including studies of leadership, management, motivation and group processes, and factors which enhance individual and institutional competence and improve life quality.

PERSONNEL MANAGEMENT SYSTEMS RESEARCH GROUP

Robert W. Stephenson, PhD, Director

Development of taxonomic systems for classifying jobs, computer assisted counseling systems and personnel data bases, assignment and career progression systems, and evaluation of individual and unit training programs.

HUMAN PERFORMANCE RESEARCH GROUP

Jerrold M. Levine, PhD, Director

Research on stress, environmental factors, information and decision processes, human abilities and skill acquisition, and psychobiological mechanisms of behavior.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Final Report	2. GOVT ACCESSION NO	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Improving Reading Comprehension: Measuring Readability		5. TYPE OF REPORT & PERIOD COVERED Final Report (15 Feb 1972 - 15 Mar 1974) 6. PERFORMING ORG. REPORT NUMBER AIR-30801-5/74-FR
7. AUTHOR(s) Ronald P. Carver		8. CONTRACT OR GRANT NUMBER(s) N00014-72-C-0240
9. PERFORMING ORGANIZATION NAME AND ADDRESS American Institutes for Research 8555 Sixteenth Street Silver Spring, Maryland 20910		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61153N RR 042-06 RR 042-06-01 NR 154-345
11. CONTROLLING OFFICE NAME AND ADDRESS Personnel and Training Research Programs Office of Naval Research (Code 458) Arlington, Va. 22217		12. REPORT DATE May 14, 1974 13. NUMBER OF PAGES 88
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Training Materials	Reading-Storage Tests	Rauding Scale of Prose Difficulty
Prose Learning	Programmed Prose	Flesch Readability Formula
Reading	Attention	Dale-Chall Readability Formula
Information Stored	Readability	Adjunct Questions
Reading Tests	Material Difficulty	Correspondence Course Material

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A standardized method, called programmed prose, has been developed which can be used to automatically convert prose training material into a form which forces trainees to read the material with at least a minimal level of comprehension. From the results of a series of experimental studies, it was concluded that programmed prose facilitates learning under the following conditions: (a) when individuals are not always highly motivated to learn, (b) when attention to the reading task wanes, (c) when the

DD FORM 1 Jan 73 1473

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. (cont'd)

reading ability level of the individual exceeds the reading difficulty level of the prose material, and (d) when the time allowed for reading exceeds the time necessary to complete the programmed prose. It was also found that programmed prose is more effective and more efficient in facilitating learning than is the study question technique that is used in correspondence course material. A new method for measuring prose difficulty, called the Rauding Scale of Prose Difficulty, was also developed and investigated so that the effect of the reading ability and reading difficulty differences could be better measured. The Rauding Scale was found to correlate higher with the actual grade level where curriculum passages were used in school than six other readability measures. It was concluded that the Rauding Scale was more valid as a measure of readability than either the Flesch or the Dale-Chall measures, i.e., the two readability formulas which have traditionally been regarded as the most valid of all formulas.

AIR 30801-5/74-FR

IMPROVING READING COMPREHENSION: MEASURING READABILITY

Ronald P. Carver
Principal Investigator

FINAL REPORT

Prepared under Contract to the
Personnel and Training Research Programs
Psychological Sciences Division
Office of Naval Research
Department of the Navy

Contract No. N00014-72-C-0240
NR No. 154-345

Approved for public release; distribution unlimited.
Reproduction in whole or in part is permitted for
any purpose of the United States Government.

American Institutes for Research
Washington Office
Human Resources Research Group

May 1974

ABSTRACT

A standardized method, called programmed prose, has been developed which can be used to automatically convert prose training material into a form which forces trainees to read the material with at least a minimal level of comprehension. From the results of a series of experimental studies, it was concluded that programmed prose facilitates learning under the following conditions: (a) when individuals are not always highly motivated to learn, (b) when attention to the reading task wanes, (c) when the reading ability level of the individual exceeds the reading difficulty level of the prose material, and (d) when the time allowed for reading exceeds the time necessary to complete the programmed prose. It was also found that programmed prose is more effective and more efficient in facilitating learning than is the study question technique that is used in correspondence course material. A new method for measuring prose difficulty, called the Rauding Scale of Prose Difficulty, was also developed and investigated so that the effect of the reading ability and reading difficulty differences could be better measured. The Rauding Scale was found to correlate higher with the actual grade level where curriculum passages were used in school than six other readability measures. It was concluded that the Rauding Scale was more valid as a measure of readability than either the Flesch or the Dale-Chall measures, i.e., the two readability formulas which have traditionally been regarded as the most valid of all formulas.

ACKNOWLEDGMENTS

The data collection and data analysis were accomplished with the assistance of Frances Cohen and Debra Stein. Especially helpful was the difficult work these two individuals performed in counting letters, words, sentences, and computing difficulty values for the readability analysis.

Without the cooperation of the Pierce City, Missouri R6 school system and the Prince Georges County, Maryland school system, this research could not have been successfully completed. The specific school personnel who contributed are noted inside the report in conjunction with each research study.

TABLE OF CONTENTS

	Page
ABSTRACT	ii
ACKNOWLEDGMENTS	iii
LIST OF TABLES	v
LIST OF FIGURES	vi
INTRODUCTION	1
MEASURING PROSE DIFFICULTY	3
Theoretical Rationale I	3
Cloze and Readability Formulas	4
The RIDE Scale	7
Study 1	8
Study 2	13
Study 3	17
Theoretical Rationale II	18
Study 4	20
Study 5	26
Study 6	34
Conclusions	39
PHASE III	40
Study 1	40
Study 2	47
PHASE IV	54
Experiment 1	55
Experiment 2	61
Experiment 3	64
Discussion	70
REFERENCES	74
APPENDIX A--Procedures for Determining RIDE Values	
APPENDIX B--Using the Rauding Scale	

LIST OF TABLES

		Page
1.	Intercorrelations among Six Estimates of Passage Difficulty and Four Estimates of Input Accuracy	12
2.	RIDE Scale Levels and the Frequency of Occurrence of Miller-Coleman Passages at each Level	13
3.	Intercorrelations among 10 Difficulty Variables for 40 Bormuth Passages	15
4.	Intercorrelations among the 8 Difficulty Variables for the 50 Bormuth Passages	18
5.	Correlations between each of the two Rauding Scale Variables and the other Difficulty Measures for the 50 Passages in Study 3	24
6.	Correlations between each of the two Rauding Scale Variables and the other Difficulty Variables for the 40 Passages in Study 2	24
7.	Percent of Curriculum Materials, by School Level, at each Rauding Scale Level	28
8.	Grade Difficulties of the Ginn Readers	30
9.	Percentage of Three Samples of Newspapers at each Rauding Scale Level	32
10.	Percentage of Graduate School Materials at each Rauding Scale Level	33
11.	Estimated Percent of Individuals who can Read and Understand Material at Various Ability-Difficulty Differences	37
12.	Descriptive Information for the Sixteen Experimental Passages	41
13.	Instructions for using the Rauding Scale of Prose Difficulty	Appendix A
14.	Determine the Grade Difficulty, G_a, of Material from the Mean of Three Rauding Scale Ratings	Appendix B

LIST OF FIGURES

		Page
1.	An example reading-input passage.	5
2.	Percentage of passages at each level difficulty, L_d, which could be read and understood by individuals at each level ability, L_a.	36
3.	Mean percent correct on the RS-tests as a function of L_a for each L_d.	44
4.	Mean percent correct on the RS-test, administered under the nonreading condition, as a function of L_a for each L_d.	45
5.	Gain in RS-test means from nonreading to reading as a function of L_a for each L_d.	46
6.	Mean scores on the Paraphrase Test for each ability level and under both the programmed prose and regular prose conditions.	50
7.	Mean scores on the RS-test for each ability level and under both the programmed prose and regular prose conditions.	51
8.	Mean percent correct for the Level 4 individuals in the regular prose, RP, programmed prose, PP, and study question, SQ, groups for both the RS-test and the MC-test, Experiment 1.	60
9.	Mean percent correct for the Level 4 individuals in the regular prose, RP, programmed prose, PP, and study question, SQ, groups for both the RS-test and the MC-test, Experiment.2	63
10.	Mean percent correct for the Level 4 individuals in the regular prose, RP, programmed prose, PP, and study question, SQ, groups for both the RS-test and the MC-test, Experiment 3 (Note: The RP group in this experiment received the test questions to study before they read and the mean score for this group is signified by a dotted line.)	67
11.	Percent gain on MC-test for the RP, PP, and SQ groups when the nonreading gain is arbitrarily considered as 0% gain and the highest score is considered as 100% gain.	68

INTRODUCTION

A standardized method has been developed which will convert prose training materials into a form which forces trainees to read the material with at least a minimal level of comprehension. The research on the method has been conducted in four phases. The Phase I and Phase II results, as well as the rationale for the research, were reported in an earlier technical report (Carver, 1973). This final report will contain the results of Phase III and Phase IV, as well as some additional research that will be explained later.

Phase I involved an extensive investigation of a new technique, called the reading-storage test, for measuring the learning that occurs during reading so that the effectiveness of programmed prose could be better assessed in Phases II, III, and IV. In six Phase I experiments, the reading-storage test was compared to two other types of tests, i.e., cloze and paraphrase. The results suggested that the completely objective, reading-storage test provides a better measure of the primary effects of reading than either the cloze test, which is developed objectively but scored subjectively, or the paraphrase test which is developed subjectively but may be scored objectively.

In the Phase II experiment, programmed prose was compared to regular prose under low and high motivation conditions. The programmed prose facilitated learning under the low motivation condition, and inhibited learning under the high motivation condition, as had been hypothesized. It appeared that programmed prose may be used to facilitate learning in reading situations wherein attention wanes.

In the two final phases of the project, the effectiveness of programmed prose was investigated further. In Phase III, the effect of material difficulty upon the effectiveness of programmed prose was investigated. In Phase IV, the conditions under which programmed prose might be expected to facilitate learning were further investigated and the facilitative effect of programmed prose was also compared to that of correspondence course material.

During the conduct of this research, it became increasingly evident that the method used for measuring the difficulty of prose material was extremely crucial. Therefore, there was additional research conducted

on measuring prose difficulty. The results of this research was used to measure and control material difficulty in Phases III and IV. Therefore, the results of this research will be presented next, i.e., prior to reporting on Phase III and Phase IV.

MEASURING PROSE DIFFICULTY

Rationale and data supporting the use of a new measure of prose difficulty will be presented. Since there are numerous readability formulas (see Klare, 1963) and since the cloze procedure has been regarded as the best measure of readability (e.g., see Bormuth, 1966; Klare, Sinaiko, & Stolurow, 1972; Rankin, 1971), it is reasonable to question the wisdom of adding another measure of prose difficulty to the already cluttered literature. The impetus for developing a new measure came from two sources. First, good measures of prose difficulty were required to conduct prose research, and there were serious inherent disadvantages associated with the existing measures, i.e., the cloze technique and readability formulas. Second, a newly developed theory of reading comprehension required accurate measures of both passage difficulty and individual ability in order to empirically test the theory.

The following series of research studies were not programmatically designed at the outset. The results of some studies stimulated other studies and the results of some studies stimulated re-analysis of previous studies. The rationale for the research will therefore be explained in connection with each successive piece of research.

Theoretical Rationale I

In the theory presented by Carver (1974a), the reading process was conceptually divided into two primary components, input and storage. The distinction is roughly similar to the distinction that is often made between reading and comprehending (see Reed, 1970). That is, there is a difference between being able to read (i.e., input) the words in a sentence, and being able to comprehend or understand (i.e., store) the complete thought in the set of words which make up a sentence. The storage of thoughts contained in sentences is directly dependent upon being able to input the individual words. Input is a necessary but not sufficient condition for storage.

When almost all of the thoughts being input are also stored, the process is called rauding. For prose rauding to occur, the accuracy of storing each successive thought contained in sentences must be high,

i.e., greater than about .75. Prose rauding theory holds that the accuracy with which an individual stores thoughts (A), is a function of the accuracy with which the thoughts are input (A_i),

$$A = \underline{f}\ (A_i). \qquad (1)$$

The accuracy with which complete thoughts are stored, A, is actually dependent upon a number of factors unique to the prose and to the individual. For example, a high school student may be able to store the thoughts contained in a passage on existentialism because his philosopher father discussed the subject with him on many occasions. This same student may be at a loss to store the thoughts contained in a passage on computers, because he was not familiar with the concepts employed. However, if the subject matter is controlled or balanced across many areas, then a general ability to store material should emerge. It is assumed that this general ability to store the thoughts would depend directly upon a general ability to input the words, i.e., as expressed in Equation 1.

It is further assumed in prose rauding theory that A_i in any particular situation is a function of the interaction between the difficulty level of the prose (L_d) and the ability level of the reader (L_a), i.e.,

$$A_i = \underline{f}\ (L_d, L_a). \qquad (2)$$

The introduction of the concept of A_i also contained a recommended measurement technique for estimating A_i. The technique is called reading-input. An example of the reading-input technique is contained in Figure 1. It is similar to the cloze technique except that instead of deleting words, alternative wrong words are added. The task for the subject is to mark the words which belong in the sentence. The alternative wrong words are selected by a completely objective method which is described in detail elsewhere (Carver, 1971; Carver, 1974d).

Cloze and Readability Formulas

As noted earlier, the cloze technique has often been regarded as a better measure of readability than any of the traditional readability formulas. This is somewhat misleading because cloze can be more

This is ☐ people ☐ our Post Office. It is ☐ a ☐ in our city. Many people ☐ post ☐ work here. There is a ☐ Post ☐ in Office in every city ☐ have ☐ in our country. And Post ☐ Here ☐ Offices in every country in ☐ the ☐ be world.

A Post Office ☐ helper ☐ a must be honest. He ☐ world ☐ must be a good worker. ☐ Must ☐ A Post Office helper handles ☐ lots ☐ must of mail. A Post ☐ sends ☐ Office helper handles lots of ☐ money ☐ lots

The Post Office sends ☐ world ☐ letters and packages, magazines, and ☐ the ☐ newspapers all over the world. ☐ It ☐ Money sends small animals and ☐ plants ☐ sends, too. It sends money ☐ for ☐ sends us. It saves money ☐ plants ☐ for us. It puts money ☐ to ☐ plants work for us, too.

Fig. 1. An example reading-input passage

accurately regarded as a method for estimating A_i, and the readability formulas can be more accurately regarded as methods for estimating L_d. Cloze scores are directly dependent upon a difficulty factor inherent in the material, such as L_d, and an ability factor inherent in the individual, such as L_a. On the other hand, the readability formulas are dependent only upon a factor inherent in the prose itself, such as L_d.

It could be argued that cloze should be as good if not better an indicator of input difficulty, A_i, as the reading-input technique. Both techniques primarily seem to reflect a knowledge of words and how they are used within sentences. Also, both techniques can be developed in an objective manner, i.e., by using an algorithm. Why then, it may be asked, develop an entirely different technique to replace cloze as an estimate of A_i?

There seem to be several advantages to using reading-input rather than cloze. Reading-input is entirely objective. It can be scored without the use of the subjective judgments necessary for scoring cloze. Reading-input is also more similar to the actual reading act in terms of task and time. The cloze task requires a subject to: (a) hypothesize various words that might fit the blank, (b) choose the best one, and (c) take the time to write the answer in the blank. The reading-input technique requires less problem solving and marking time in relation to reading time. As a result, the reading-input technique requires far less time and mental energy expenditure on the part of the subject. The major disadvantage of the reading-input technique is the time and effort required by the experimenter in choosing the incorrect alternative words. Yet, much of this lost time is saved during the scoring process since reading-input may be scored with a regular punched scoring key while every replaced word must be subjectively scored on the cloze test. It is anticipated that the effort involved, at present, in developing reading-input material will be eliminated in the future by computer techniques. A properly programmed computer could be presented prose passages and reading-input materials could be produced automatically. Both reading-input and cloze could be used as indicants of A_i, but reading-input seems to have many advantages over cloze as the preferred way of estimating A_i.

Cloze is also regarded as an indicant of difficulty, L_d, as well as A_i, but this use of cloze is only valid in certain situations. If there were a standard group of individuals who could be used to fill out cloze blanks on passages, then cloze would provide an indicant of A_i since L_a would always be constant in Equation 2. In most situations, however, cloze is a rubber yardstick and cannot be used as an absolute index of input difficulty. This is because the cloze index of difficulty for a particular prose passage may be .70 when the subjects completing the cloze task are college students, i.e., L_a is high, whereas the index for the same passage may be .30 when the subject are second graders, i.e., L_a is low.

Cloze difficulty values for a passage vary, but a difficulty value obtained by a traditional readability formula does not vary, i.e., there is only one difficulty value for any particular formula. Rather than developing a new method for estimating L_d, one of the existing readability formulas could have been chosen. However, almost all of these existing methods have the following inherent disadvantage: certain properties of passages are isolated and measured and then an empirical formula is developed by correlating predictor variables with a criterion. The problem with this procedure is that the validity of the resulting formula is directly dependent upon the validity of the criterion used in the initial research. Therefore, it was decided that a new method for estimating L_d should be developed which would provide a scale that was not directly dependent upon a derived criterion. The following section describes the selected variable.

The RIDE Scale

Klare (1968) presented evidence that the frequency of occurrence of words affects prose difficulty and that words get shorter as they are used more frequently. Thus, some measure of word length should provide an indicant of prose difficulty. Bormuth (1966) interpreted his data as indicating that a count of the number of letters provided a better measure of word length than a count of syllables. Therefore, the variable chosen for estimating difficulty, L_d, for a passage was the

average number of letters-per-word. The average number of letters-per-word is taken to provide a Reading-Input Difficulty Estimate, i.e., the RIDE Scale.

Empirical evidence supporting average word length as an estimate of prose difficulty, i.e., the RIDE Scale, already exists in the literature (see Bormuth, 1968). Bormuth (1969) analyzed 169 readability variables with respect to their correlations with the cloze measure of prose difficulty. This study involved approximately 2,600 students in grades 4-12 and 330, 100-word passages. Out of the 169 variables he studied, the average number of letters-per-word was one of the two highest correlates of the cloze difficulty ($r = -.721$); the Dale-Chall measure correlated .006 higher. In another study, Coleman (1971) studied 32 predictor variables and found that the number of letters-per-word correlated highest ($r = -.89$) with the cloze scores on the 36 Miller-Coleman paragraphs (see Aquino, 1969). It appears that when individual differences in ability are held constant, one of the highest correlates of input accuracy is the average number of letters-per-word. That is, using the cloze scores as an estimate of A_i, the average number of letters-per-word appears to be a highly valid estimate of L_d when L_a is held constant.

When cloze is used as an estimate of input accuracy, A_i, then the RIDE Scale appears to provide a valid estimate of difficulty, L_d. Yet, the best way of directly evaluating the validity of the RIDE Scale would be to investigate its correlation with input accuracy, A_i, as estimated by the reading-input technique. Study 1, which follows, was this type of an investigation.

Study 1

Introduction

The primary purpose of this study was to determine the accuracy with which reading-input scores could be predicted from the RIDE Scale when individual differences in ability were controlled. Secondary purposes of the study were: (a) to compare the RIDE Scale to other estimates of L_d, i.e., various readability formulas, and (b) to compare the reading-input technique to the cloze technique as an estimate of A_i.

Method

Subjects. The Ss were 180 students in a small midwestern school system.[1]

Passages. The 36 Miller-Coleman passages (see Aquino, 1969) were used. These are 150-word passages ranging from very easy to very difficult material. The reading-input technique was applied to these passages using the procedures described elsewhere (Carver, 1974b). There was one item for each five words ($N_r = 5$) giving a total of 30 items per passage. The number of incorrect alternatives per item was one ($N_a = 1$).[2] The reading-input technique was applied to each passage five different times to produce five different forms. Each form involved one of the five possible item positions. Words 1, 6, 11, ... 146, constituted the item position for Form 1 ($X_i = 1$) while words 2, 7, 12, ..., 147 constituted the item position for Form 2 ($X_i = 2$), and so on, for the five possible forms. There were a total of 180 different reading-input passages, five for each 36 original Miller-Coleman passages.

Procedure. Each class was told at the outset that they were participating in a scientific experiment and were asked to do their best. Then, the test booklets were distributed. Each booklet contained four reading-input passages. On the cover of the booklet was an example reading-input passage and directions. The example and the directions were read to the Ss. Then the Ss were asked to complete the practice passage which was also on the cover. The E checked each S's practice passage individually to assure that the directions had been understood. In a few cases, extra help in understanding the task was required, and this help was administered on an individual basis.

---

[1] Pierce City, Missouri R6 School District--Mr. J. D. Smith, Superintendent; Mr. Earle Staponski, Elementary Principal; Mr. Donald Trotter, Secondary Principal.

[2] The other parameters used to develop the reading-input passages were as follows: $N_p = 25$, $N_b = 4$, $X_j = 3$, and $N_c = 4$. An explanation of the meaning of these parameters is given with the procedures for developing reading-input passages.

The Ss were told to take as much time as they wanted to complete the tasks. They were also told that their booklets would be checked by E when they were finished to make sure they had marked an answer to every item. The time to complete the four passages varied from about 4-15 minutes.

Design. The 36 Miller-Coleman passages were rank ordered with respect to their RIDE Scale values, and then divided into four hierarchical blocks, A, B, C, and D, of nine passages each. The Ss 1-9 received Form 1 of each reading-input passage. Ss 10-18 received Form 2, and so on up to Ss 37-45 who received Form 5. As noted earlier, each of these 45 Ss completed four reading-input passages. Within each set of nine Ss, the first S received the lowest ranked passage in each of blocks, A, B, C and D, the second S received the second lowest ranked passage in each of the blocks A, B, C and D, and so on until the ninth S received the ninth ranked passage in each of the blocks A, B, C, and D.

There were 24 possible presentation orders for the four passages presented to each S. Each S in the first group of 24 Ss received one of the 24 possible presentation orders. Since there were 45 Ss, the second set of 24 contained only 21 Ss.

This design was replicated for each of four grades in school, i.e., grades 3, 6, 9, and 12. All students who were present that day in each grade were tested. The first 45 Ss receiving a test booklet constituted the experimental group. Specially coded booklets exactly like the experimental groups booklets were prepared for the overage in each class, but these booklets were not included in the study.

The above design procedures assured that: (a) each S would receive one passage from each quartile of the range of difficulty represented by the 36 Miller-Coleman passages, (b) each set of 45 Ss would receive all five forms of the reading-input passages, (c) any order of presentation effects would be almost perfectly distributed among all 36 passages, and (d) a wide range of ability would be represented.

RIDE Scale. The RIDE value for a passage is the average number of letters-per-word using certain decision rules designed to control for atypical or unrepresentative prose segments, e.g., numbers and proper nouns. These rules are presented in Appendix A.

Readability Formulas. Five readability formulas were evaluated as L_d estimates--Flesch Reading Ease (Flesch, 1948), Dale-Chall[3] (Dale and Chall, 1948), Automated Readability Index (Smith and Kincaid, 1970), SMOG (McLaughlin, 1969), and Fry (Fry, 1968). The uncorrected Dale-Chall values and the Flesch Reading Ease Scores were the variables that were analyzed. The Flesch and Dale-Chall were used because the research reviewed by Klare (1963) seemed to indicate that of the more than 30 available formulas, these two were the most valid. The Automated Readability Index (ARI), SMOG, and Fry represent more recent formulas.

Criterion Variables. The most important criterion for estimating A_i was the reading-input variable. The score on each reading-input passage was adjusted for guessing using a "rights minus wrongs" formula. Each of the 36 Miller-Coleman passages had a single reading-input score which was the median of 20 scores. The median of the 20 scores represented 20 different individuals, four school grade levels, and five reading-input forms.

The second most important criterion variable was the Miller-Coleman cloze score for each of the 36 passages, as given in the article by Aquino (1969).

Two other criterion variables were taken from the research of Aquino. The Recall variable was a verbatim type recall with each correct noun, verb, adjective, and adverb scored as one. The Judged Difficulty variable was the average rank of all subjects, with each subject being asked to rank the 36 paragraphs from 1 to 36 with respect to difficulty.

Results and Discussion

The means, standard deviations, and intercorrelations among the ten variables are presented in Table 1.

Of the six estimates of L_d, the highest correlate of reading-input was the Dale-Chall variable, -.90. However, the second highest correlation was .87, representing both the RIDE and Flesch variables.

Of these six variables, the highest correlate of cloze was the Flesch, .90. The second highest correlate was the Dale-Chall variable at -.89 and third highest was the RIDE variable at -.88.

---

[3]Walter H. MacGinitie provided the Dale-Chall grade level score for each of the 36 passages.

Table 1

Intercorrelations among Six Estimates of Passage Difficulty and Four Estimates of Input Accuracy

Variables		2	3	4	5	6	7	8	9	10	Mean	S.D.
Estimates of L_d												
RIDE	1	-.94	.93	.83	.84	.79	-.85	.87	-.88	-.87	4.3	.5
Flesch	2		-.96	-.92	-.90	-.95	.87	-.91	.90	.87	69.4	21.5
Dale-Chall	3			.89	.89	.88	-.86	.90	-.89	-.90	6.4	2.0
ARI	4				.89	.90	-.80	.87	-.84	-.81	8.7	4.4
SMOG	5					.87	-.72	.84	-.80	-.80	8.9	3.6
FRY (N = 31)	6						-.75	.85	-.82	-.71	6.8	2.8
Estimates of A_i												
Recall	7							-.89	.91	.82	33.0	11.4
Judged Diff.	8								-.94	-.87	18.5	8.7
Cloze	9									.87	.54	.14
Reading-Input	10										26.0	3.2

For this sample of passages, there seems to be little difference between the predictive validity of Dale-Chall, Flesch, and RIDE. Each correlates about the same with either reading-input or cloze. The intercorrelations among these three estimates of L_d are .94, .93, and .96. It appears that the RIDE Scale, the simplest index of the three to calculate, is just as valid an estimate of L_d as the other two more established formulas.

Given that these three variables are equally good estimates of input difficulty, then their correlations with cloze and reading-input should indicate which of these two variables represents the best index of input accuracy, A_i. The six correlations in question range only from .87 to .90, thus suggesting that cloze and reading-input are equally valid as indicants of A_i.

One of the most obvious aspects of all the data in Table 1 is its homogeneity. The four criterion estimates of A_i all correlate about equally high with the estimates of L_d. The 45 correlations range only from .71 to .96, and more than half of them are between .85 and .90. The lowest correlation, .71, involves the Fry formula, but this is not surprising since the Fry variable was restricted in the upper range to Grade 12. The Fry is based upon the Flesch formula, and it correlated

-.95 with Flesch. The Fry, ARI, and SMOG indexes all appear to provide valid estimates of input difficulty, but on the whole they appear to be slightly less valid than the Dale-Chall, Flesch, and RIDE Scale.

In summary, all of the potential estimates of A_i and L_d appear to be approximately equal in validity. There appears to be no evidence to suggest that any estimate of L_d is significantly more valid than the RIDE Scale. There appears to be no evidence that any estimate of A_i is significantly more valid than reading-input.

Study 2

Introduction

A further analysis of the Miller-Coleman passages using the RIDE Scale provided a troublesome finding. The RIDE Scale was arbitrarily divided into five levels in increments of .5 words and the number of passages at each level was determined. The results of this analysis are provided in Table 2. Notice that there are less passages at each successively higher RIDE level. This suggested that the Miller-Coleman passages may not be representative of the population of passages expected in the real world. Therefore, it seemed desirable to study another set of passages to see if the validity of the RIDE Scale could be replicated.

Table 2

RIDE Scale Levels and the Frequency of Occurrence of Miller-Coleman Passages at each Level

RIDE Scale (letters per word)	Level	No. Miller-Coleman Passages
⟶ 4.0	1	14
4.1 ⟶ 4.5	2	12
4.6 ⟶ 5.0	3	6
5.1 ⟶ 5.5	4	3
5.6 ⟶	5	1

In connection with the development of a standardized reading test to measure L_a (Carver, 1974), it became possible to further evaluate the RIDE Scale using another set of experimental passages.

Method

Passages. The passages were sampled from Bormuth's (1969) 330, 100-word passages. These 330 passages were sampled from curriculum materials at five school levels--Level 1, Grades 1-3; Level 2, Grades 4-6; Level 3, Grades 7-9; Level 4, Grades 10-12; Level 5, College. The RIDE Scale values for these 330 passages were determined, and then the passages were sorted into the five RIDE Levels. Ten passages were then sampled from the population at each of the first four RIDE Levels, giving a total of 40 passages. There were very few passages at RIDE Level 5 so that this level was not sampled.

The 10 passages at each level were selected by a stratified sampling procedure to represent each level. At each RIDE Level, the population of passages was divided into the five Bormuth Levels, and the number of passages selected from each of these Bormuth Levels reflected the relative frequencies of these levels. The Cloze scores, as reported by Bormuth, within each Bormuth Level and within each RIDE Level were used to select the particular passages. The passage or passages which were at or nearest to the cloze median for each of the Bormuth Levels within each RIDE Level were the passages selected.

Difficulty Measures. The same five formulas were evaluated in Study 2 as were in Study 1--Flesch Reading Ease, Dale-Chall, ARI, SMOG, and Fry. Two other difficulty variables were the Cloze scores for each passage and the Bormuth Level of each passage.

An additional variable that was unique to this study was a Teacher Rating variable. Teachers were to rate each student in their class with respect to whether the student could or could not read and understand each of the 40 different passages. The 40 passages were randomly ordered and given a new identification number from 1 to 40 reflecting this order. The instructions to the teachers were: (a) read passage No. 1, (b) decide whether each one of your students can or cannot read and understand at least 75% of the passage, (c) mark a box beside each student's name reflecting this judgment, and (d) repeat this procedure

for all 40 passages. The teachers were not rating the difficulty of the passages directly, but were rating the ability of the individuals. By summing over all individuals for each passage, the proportion of the total population in each school which could read and understand each passage was calculated. This proportion was the Teacher Rating variable for each passage.

There were two Teacher Rating variables. The Teacher Rating$_1$ variable was derived from four teachers rating 99 students in school grades 2-6.[4] The Teacher Rating$_2$ variable was derived from 11 teachers rating 239 students in school grades 2-6.[5]

Results

Table 3 contains the means, standard deviations, and intercorrelations for all of the ten variables.

Table 3

Intercorrelations among 10 Difficulty Variables for 40 Bormuth Passages

Variables		2	3	4	5	6	7	8	9	10	Mean	S.D.
RIDE	1	-.90	.87	.84	.88	.89	-.89	-.88	-.81	.85	4.50	0.57
Flesch	2		-.90	-.92	-.93	-.91	.87	.85	.80	-.83	63.1	26.6
Dale-Chall	3			.82	.88	.87	-.89	-.87	-.86	.86	7.1	1.9
ARI	4				.91	.86	-.82	-.75	-.69	.76	9.0	5.6
SMOG	5					.90	-.85	-.88	-.82	.78	10.2	3.8
FRY	6						-.85	-.85	-.81	.83	9.3	4.8
Cloze	7							.92	.91	-.93	.369	.120
Teacher Rating$_1$	8								.96	-.89	.477	.336
Teacher Rating$_2$	9									-.89	.457	.310
Bormuth Level	10										3.1	1.5

[4]The cooperation of the teaching staff of Calvary Lutheran School, Silver Spring, Maryland is gratefully acknowledged--Mr. Ellsworth Kierbs, Principal.

[5]The help of the teachers in the Pierce City, Missouri School System, noted earlier, was very much appreciated.

For the most part, the correlations in Table 3 replicate the corresponding correlations in Table 1. The intercorrelations among RIDE, Flesch, Dale-Chall, ARI, SMOG, Fry and Cloze were all approximately the same between the two sets of passages. The Fry correlations changed somewhat due to the fact that the Fry values in this analysis above Grade 12 were interpolated to give values for all 40 variables. This had the effect of lowering the correlation between Fry and Flesch, but in general raised the correlations between Fry and the other variables.

The Teacher Rating variable was very reliable in that Teacher Rating_1 correlated .96 with Teacher Rating_2. This would seem to indicate that this could be considered as an excellent criterion variable for evaluating the other difficulty measures. It may be noted that the highest correlate of the two Teacher Rating variables was Cloze, .92 and .91, and the second highest correlate was the Bormuth Level, -.89 and -.89.

Cloze, Teacher Rating, and Bormuth Level may be considered as criterion variables for comparing RIDE with the five readability formulas. Using Cloze as the criterion, RIDE and Dale-Chall correlated the highest, -.89, while ARI correlated the lowest, -.82. Using Teacher Rating_1 as the criterion, RIDE and SMOG correlated the highest, .88, while ARI again correlated the lowest, -.75. Using Teacher Rating_2 as the criterion, Dale-Chall correlated the highest, .86, while ARI again correlated the lowest, -.69. For the Bormuth Level, RIDE correlated second highest, .85, only .01 below Dale-Chall, and ARI again correlated the lowest, .76.

Discussion

These data suggest that the RIDE Scale is just as valid as any of the readability formulas, since it was one of the highest correlates of the four criterion variables. However, there is also something disturbing about these data. The passages were selected on the basis of three of the four criterion variables--RIDE, Cloze, and Bormuth Level. Thus, the intercorrelations among these variables are not complete free to vary, but were influenced to an unknown degree by the selection

process itself. Therefore, it seemed desirable to withhold any final conclusions about the validity of the RIDE Scale, since there were peculiarities about both the Study 1 and Study 2 samples of passages.

Study 3

Introduction

It was decided that another set of passages should be sampled to evaluate the RIDE Scale further. The Bormuth Level from which the curriculum passages were selected seemed to provide an excellent criterion variable for evaluating the validity of the various measures. Therefore, another study was conducted by again sampling from the Bormuth passages but this time the only restriction was that an equal number of passages be sampled from each Bormuth Level.

Method

Passages. These were 10 passages randomly sampled from the population of Bormuth passages at each of the five levels, giving a total of 50 passages.

Difficulty Measures. There were eight difficulty measures in Study 3, the same as Study 2 except for the lack of the two Teacher Rating variables.

Results

The intercorrelations, means, and standard deviations of the eight variables are presented in Table 4. The Cloze and Dale-Chall variables correlated highest with the Bormuth Level, -.82 and .82, while the Fry and Flesch variables correlated only .01 lower, .81 and -.81. The ARI and SMOG correlated slightly lower at .76 and .79 respectively, while the RIDE Scale correlated lowest of all at .70.

Using Cloze as the criterion, the RIDE Scale was the lowest correlate at -.77 and the Dale-Chall was the highest correlate at .84. The second and third highest correlates of Cloze were the Fry and Flesch at -.82 and .81, respectively.

Table 4

Intercorrelations among the 8 Difficulty Variables for the 50 Bormuth Passages

Variables		2	3	4	5	6	7	8	Mean	S.D.
RIDE	1	-.90	.86	.80	.86	.87	-.77	.70	4.55	.47
Flesch	2		-.92	-.92	-.97	-.95	.81	-.81	60.6	25.7
Dale-Chall	3			.83	.86	.85	-.84	.82	7.18	2.17
ARI	4				.91	.83	-.80	.76	10.1	6.3
SMOG	5					.89	-.79	.79	10.7	4.5
FRY	6						-.82	.81	9.2	4.7
Cloze	7							-.82	.374	.133
Bormuth Level	8								3.00	1.43

Discussion

These data make it difficult to contend that the RIDE Scale is as valid as the Flesch and Dale-Chall. It seems that there is very little difference between any of the six variables--RIDE, Flesch, Dale-Chall, ARI, SMOG, and Fry--but that the Flesch and Dale-Chall seem to be consistently better than the others. These findings seem to further support the contention of Klare (1963) that the Flesch and Dale-Chall are the most valid readability formulas available.

Theoretical Rationale II

The data seemed to suggest that the best available indicator of difficulty level, L_d, is either the Flesch or Dale-Chall. This was somewhat difficult to accept since several passages had been encountered in the three studies which were assigned difficulty values that were counter intuitive. For example, a passage on what seemed to be a conceptually complex topic might be assigned a Grade 5 difficulty because the sentences were short and a small percent of the words were not on the Dale-Chall list. Syllable counts and sentence length also provided some Flesch difficulty values that did not seem appropriate.

It seemed desirable to develop a more thorough rationale for what it is that we desire to measure before proceeding further with developing and evaluating the empirical measures themselves. The rationale that was developed will be presented in the following paragraphs.

The degree to which individuals have difficulty understanding passages would seem to depend upon at least three major factors--vocabulary,

ideas, and style. With respect to vocabulary, some passages contain mostly common words and some passages contain mostly uncommon words. The Dale-Chall reflects this factor by use of the Dale-Chall list of frequently used words; the Flesch reflects this factor by measuring the average syllable length of the words; and the RIDE Scale reflects this factor by the average length of the words in letters. With respect to style, some passages are written using short, simple sentences, while some passages are written using long, complex clauses, phrases, and sentences. The Flesch and Dale-Chall reflect this factor by measuring average sentence length in words. However, sentence length does not measure the grammatical complexity of the sentences, it is only a correlate of grammatical complexity. Indeed, the results of Studies 1-3 indicate that average sentence length does not add much to the prediction of difficulty since the Dale-Chall and Flesch, which use sentence length, were not much more valid than the RIDE Scale, which has only a vocabulary or word factor. Furthermore, neither sentence length nor sentence complexity reflect the type of style difficulty that arises from inter-sentence dependency, e.g., poor paragraph construction and ambiguous use of anaphora. With respect to idea difficulty, some passages are written using simple ideas or concepts while other passages are written using complex ideas or concepts. For example, the concepts of "over" and "under" are much easier for most individuals to understand than the concepts of "voltage" and "amperage." None of the potential indicators of L_d seem to directly reflect idea or concept difficulty.

Since none of the potential indicators of L_d directly reflect idea difficulty, none are candidates for use in rewriting prose so as to make it easier to understand. Klare (1963) has warned that readability formulas are not designed to be used in rewriting prose. It would be quite easy for someone to revise difficult material by chopping sentences in half and inserting a large number of short words, e.g., "of," "the," "and." Although the Flesch and Dale-Chall values that resulted after such a revision would suggest that the material was less difficult, in fact, the material would likely be much more difficult to read and understand. The problem involved in using either the Flesch or

Dale-Chall technique might be better illustrated using two hypothetical examples. If the Dale-Chall or Flesch difficulty values of a passage were determined, and then the order of the sentences within the passage were randomly rearranged, the new difficulty values would be exactly the same as the original values. Thus, this modified form of the passage would be much more difficult to read and understand, but this change would not be reflected in the Dale-Chall or Flesch values. For another hypothetical example, consider a passage where the order of the words within the sentences were scrambled, i.e., rearranged in random order. This modified form of the passage would have exactly the same Dale-Chall and Flesch values even though the modified version would be impossible to read and understand, i.e., it would be gibberish. The Dale-Chall and Flesch could not discriminate between gibberish and regular prose, thus indicating the limitations of these measures.

What is needed is a measure of difficulty which would be sensitive to all of these factors which tend to make a passage difficult or easy to understand. Thus, it may be said that what is needed is a more valid, objective method for measuring the difficulty associated with understanding a passage, i.e., a more valid method for measuring L_d. With the above criterion as a guideline, the Rauding Scale of Prose Difficulty was developed. It will be described in the next study.

Study 4

Introduction

In Study 2, described earlier, the Teacher Ratings of passage difficulty were extremely reliable, and presumably valid, even though they involved a different set of teachers in a different part of the country who were not rating passages at all but were rating the ability of their students. In a different area of research, Carroll (1971) found that professional lexicographers were extremely reliable in estimating the frequency of usage of words, and it was his opinion that these professional judgments were probably more valid than the extensive empirical data that had been collected on word frequency. Thus, it seemed possible that teachers, or other qualified experts, might be

employed to produce objective ratings of passage difficulties which, in turn, might be more valid than the Dale-Chall or Flesch formulas.

Such a method was developed; it is called the Rauding Scale of Prose Difficulty. The Rauding Scale consists of an anchor set of six passages ostensibly representing Grades 2, 5, 8, 11, 14, and 17. A qualified expert reads the passage to be rated and then decides where it fits on the Rauding Scale using the six passages as anchor passages. For example, if a qualified expert decided that a passage was much more difficult than the Grade 5 passage but a little less difficult than the Grade 8 passage, the expert might assign the passage a Grade 7 difficulty rating.

In order to become qualified to use the Rauding Scale, an individual must pass the Rauding Scale Qualification Test (Carver, 1974b). The test simply requires that an examinee accurately rate five passages using the Rauding Scale. The Rauding Scale grade difficulty of a passage is determined by: (a) getting the average of the judgments of three qualified raters, and (b) entering a table with this average value to determine the final Rauding Scale grade difficulty value.

In making Rauding Scale judgments, the qualified expert is asked to: (a) assign a grade to a passage which designates the grade where the rater thinks the average reader should be able to read and understand most of the passage, and (b) use the six Rauding Scale passages as examples of what to expect an average reader to be able to read and understand at each grade. The rater is asked to make ratings about reading and understanding passages so the scale has been called the Rauding Scale because the word rauding has been defined as encompassing both reading and understanding (see Carver, 1974a).

Appendix B contains a complete listing of the procedures required to determine the grade difficulty of a passage using the Rauding Scale. The six anchor passages used for the Rauding Scale in particular were selected after a long series of pilot studies wherein the passages were ranked and rated by several raters in several ways. For Grade 2, the trial anchor passages were always selected from passages in Bormuth Level 1, i.e., grades 1-3. For Grade 5, the trial anchor passages were always selected from passages in Bormuth Level 2, i.e., Grades 4-6. This method for selecting trial anchor passages was similar for each grade up to the Grade 14 trial passages which were

always selected from the Bormuth College Level. The Grade 17 trial anchor passages were selected from a set of 30 graduate school level passages whose characteristics will be explained in detail later, i.e., in Study 5. When a passage did not provide ordinal rankings in the manner expected, a new candidate for that grade level was substituted, and more data collected. The final six anchor passages seemed to provide relatively reliable ordinal rankings. Thus, the final anchor passages were curriculum passages which were used in school at or near the particular grade that they were supposed to represent on the Rauding Scale and pilot data indicated that raters seemed to agree that each higher grade passage was more difficult than the lower grade passages. The ratings derived from using the Rauding Scale anchor passages have been further scaled so that they more precisely represent their particular grade difficulties. This final scaling is explained at the end of Appendix B.

The Rauding Scale was applied to the Study 2 and Study 3 passages in order to evaluate the reliability and validity of the scale.

Method

Selection of Qualified Experts. The Rauding Scale Qualification Test was administered to 60 students enrolled in three graduate reading courses at the University of Maryland. Out of this group, 10 passed the test. Six of these 10 qualified raters were hired to rate the grade difficulty of the Study 2 and Study 3 passages.

Procedures. The first three qualified experts constituted Group A and the second constituted Group B. Each member in each group was given the instructions contained in Appendix B on an individual basis. Each rater was further asked to read each passage for one minute before assigning a grade value to the passage. To assure that no one hurried through the task, the S was instructed to activate a timer prior to reading each new passage. When the timer was activated, a light went on and when the 60 seconds were up the light went off automatically. The rater was instructed to take as much time as was desired to rate each passage, but if the rater finished prior to the 60 sec. minimum, he or she was to wait until the 60 sec. had expired before starting on the

next passage. The set of 40 passages from Study 2 was administered first, and the set of 50 passages from Study 3 was administered second. The order of administration of the passages within each set of 40 or 50 was randomized, i.e., the passages were shuffled for each rater to control for order effects.

Results

Table 5 contains the correlations between each of the two Rauding Scale values and all of the other difficulty variables in Study 3. The correlations between the criterion variable, i.e., the Bormuth Level, and all of the other variables from Table 4 have been included again in this table so as to facilitate comparisons.

The correlation between the two Rauding Scale values, i.e., Group A and Group B was extremely high, .97, signifying high realiability for the Rauding Scale. The size of the sample for the correlations in Table 5 was actually 49, not 50, because one of the 50 randomly selected passages happened to be an anchor passage on the Rauding Scale.

The highest correlates of the Bormuth Level were the two Rauding Scale variables, .85 and .85. The Dale-Chall and Cloze were the two next highest correlates at .82, and the Flesch and Fry both correlated .81. Next was the ARI at .78, SMOG at .76, and last was the RIDE Scale at .69. If the Rauding Scale was considered to be the criterion variable, instead of the Bormuth Level, the validity ranking of the formulas would be approximately the same.

Table 6 contains the corresponding correlations for Study 2. In this study the Cloze and the RIDE Scale correlated highly with the Bormuth Level, -.93 and .86 respectively, but these correlations should not be interpreted as representative because the passages were selected on the basis of their RIDE, Cloze, and Bormuth Level values. The two Rauding Scale variables correlated highly with the Bormuth Level, .90 and .87; these two correlations were higher than both the Flesch, -.83, and the Dale-Chall, .85, as well as the other three readability formulas. Thus, with the Bormuth Level as the criterion, the Rauding Scale correlated higher than any of the six readability variables; the 40 passage sample thus replicates these same results for the 50 passage sample.

Table 5

Correlations between each of the two Rauding Scale Variables and the other Difficulty Measures for the 50 Passages in Study 3

Variables	Bormuth Level	Rauding Scale	
		Group A	Group B
RIDE	.70	.75	.74
Flesch	-.81	-.80	-.79
Dale-Chall	.82	.82	.80
ARI	.76	.71	.71
SMOG	.79	.80	.78
Fry	.81	.79	.78
Cloze	-.82	-.82	-.82
Rauding Scale, Group A	.85	--	.97
Rauding Scale, Group B	.85	.97	--
Mean		6.9	6.3
S.D.		4.7	4.4

Table 6

Correlations between each of the two Rauding Scale Variables and the other Difficulty Variables for the 40 Passages in Study 2

	Bormuth Level	Rauding Scale	
		Group A	Group B
RIDE	.85	.79	.84
Flesch	-.83	-.77	-.80
Dale-Chall	.86	.83	.82
ARI	.76	.64	.67
SMOG	.78	.78	.83
Fry	.83	.78	.79
Cloze	-.93	-.87	-.86
Teacher Rating$_1$	-.89	-.94	-.96
Teacher Rating$_2$	-.89	-.92	-.92
Bormuth Level	--	.90	.87
Rauding Scale, Group A	.90	--	.94
Rauding Scale, Group B	.87	.94	--
Mean		7.4	7.4
S.D.		4.6	4.8

It may be noted also that the four intercorrelations among the Teacher Rating variables and the Rauding Scale variables, -.94, -.96, -.92 and -.92, were approximately equal to the reliability estimate of the Rauding Scale itself for this set of passages, i.e., .94.

Discussion

The two reliability estimates for the Rauding Scale are extremely high, i.e., .97 and .94, thus indicating that the Rauding Scale can be made relatively objective and reliable by using the collective wisdom of experts. The Rauding Scale also correlated higher with the actual grade level wherein the passages were used in school, i.e., the Bormuth Level variable, than did the Flesch or the Dale-Chall. The level at which the passages were used as curriculum materials is probably as valid an empirical variable as is available to compare these estimates of L_d. Therefore, it appears reasonable to conclude that the Rauding Scale is more valid than the Flesch and the Dale-Chall.

The extremely high correlations between the Teacher Rating variables in Study 2 and the Rauding Scale variables also provides strong evidence for the high validity of the Rauding Scale. On the surface, it may appear that this is an artificially high relationship, but it is not. The Teacher Rating variables did not involve the Rauding Scale nor were the teachers rating passage difficulty at all. The teachers were rating their students' ability to read and understand each passage, and it was a statistical manipulation, unknown to the teachers, which translated these ratings of student ability into ratings of passage difficulty. When these two sets of teachers in schools over 1,000 miles apart can produce passage difficulty data which correlate .96, and when these data correlate .92 to .96 with the Rauding Scale variables, it seems reasonable to conclude that the Rauding Scale is extremely valid. The Rauding Scale appears to be about as valid as it is reliable, i.e., about .94 to .97, i.e., its validity is only limited by its reliability.

It should not be too surprising that the Rauding Scale is more valid than the Dale-Chall or Flesch for estimating the relative difficulty with which an individual can read and understand a passage. Neither the Dale-Chall or the Flesch attempt to directly reflect the

difficulty of the ideas or concepts in a passage, one of the most important factors affecting understanding, whereas the Rauding Scale does attempt to directly reflect this factor. Because the Rauding Scale does directly reflect the difficulty of understanding, it is the only known estimate of L_d which is appropriate for writing or revising material. Whereas the difficulty estimates produced by the traditional readability formulas, such as the Flesch or Dale-Chall can be spuriously lowered when difficulty of understanding is actually raised, such data should be nearly impossible to produce using the Rauding Scale. When a passage is revised so as to contain choppy sentences and inappropriately inserted little words, the Flesch and Dale-Chall formulas may suggest that it is much easier to read and understand whereas the Rauding Scale is likely to more validly indicate that the material is much harder to read and understand.

In summary, the Rauding Scale appears to be extremely reliable and valid, and more valid than the Flesch or Dale-Chall in estimating the relative difficulty that many individuals will encounter when attempting to read and understand a passage. The Rauding Scale also appears to be the only available estimate of L_d which is inherently appropriate for use in writing and revising material.

Study 5

Introduction

The Rauding Scale provides difficulty estimates in grade units from Grade 1 to Grade 18, and this variable which reflects difficulty in grade units is symbolized by G_d. When the Rauding Scale is divided into levels, this variable is symbolized as L_d. The six levels of L_d include three grades of G_d at each level, i.e., $L_d = 1$ when $G_d = 1$, 2, or 3; $L_d = 2$ when $G_d = 4$, 5, or 6, etc. The Rauding Scale appears to be highly reliable and valid in estimating relative passage difficulties of curriculum materials used in school, i.e., Grade 1 to College. However, the validity of the Rauding Scale as an absolute indicator of grades or levels of difficulty remained to be investigated. For example, it is not likely but it is possible that the Rauding Scale would rank the

relative difficulty of a set of passages very reliably and validly but assign first grade material a Grade 10 difficulty value. The absolute validity of the grade difficulty values and the level difficulty values was investigated by applying the Rauding Scale to: (a) curriculum materials used in school, (b) basal reading materials, (c) newspapers, and (d) graduate school materials. These data will allow an evaluation of the degree to which the Rauding Scale produces grade and level values which are commensurate with what would be expected from a knowledge of the materials themselves.

Curriculum Materials

Population and Sample. The population of interest was the curriculum materials used in school from Grade 1 to College. As a sample of this population, the 330, 100-word passages sampled by Bormuth (1969) was used. As noted earlier, this sample is divided into five levels--Grades 1-3, Grades 4-6, Grades 7-9, Grades 10-12, and College. There are 33 passages in each of 10 subject matter areas, e.g., biology, mathematics, and current events. These five levels of Bormuth correspond to the first five Rauding Scale levels except that Level 5 for Bormuth was defined simply as College whereas Level 5 on the Rauding Scale is defined as Grades 13-15.

Rauding Scale. The Rauding Scale values for the 330 passages were produced by the same Group A raters using the same procedures as were employed in Study 4. These ratings took place over a period of at least three days, with no more than three hours of rating taking place on any one day. The 330 passages were randomly divided into eight sets of 40 passages each and one set of 10 passages. The nine sets were presented in a different randomized order to each rater, and the order within each set was randomized, i.e., the 40 passages were shuffled for each rater.

Results and Discussion. Table 7 contains the percent of the passages at each Bormuth Level which are at each Rauding Scale Level, i.e., L_d. It may be noted that more than half of the passages at each Bormuth Level are at the corresponding Rauding Scale level. For Bormuth Level 1 (Grades 1-3), 76 percent of these passages are at Rauding Scale Level 1 (Grades 1-3). The corresponding percentages for the remaining

levels are as follows: Level 2, 56 percent; Level 3, 52 percent; Level 4, 56 percent; and Level 5, 70 percent. It should not be surprising that the Rauding Scale levels match the Bormuth Levels so closely since the Rauding Scale was anchored to these Bormuth Levels (see Appendix B). What is more informative about these data is the degree of variability around the expected level. It would not be expected that all of the curriculum materials used in Grades 7-9 (Bormuth Level 3), for example, would be at $L_d = 3$. Some individuals in Grades 7-9 would be capable of reading and understanding material at a higher level so it would seem reasonable that some of the curriculum materials would be at a higher level. Similarly, some individuals in Grades 7-9 would be incapable of reading and understanding material at Grades 7-9 difficulty, so it would seem reasonable that some of the curriculum materials in Grades 7-9 would be at a lower level of difficulty. However, it would not seem to be reasonable to find a large percentage of the curriculum materials at grade levels of difficulty that are a great deal higher or a great deal lower than the level at which they are being used. Indeed, this appears to be the case with the data in Table 7, since an average of only about three percent of the curriculum materials are more than one level higher in difficulty than the level at which they are used in school. The corresponding average is five percent for the percent of curriculum materials that are more than one level below the level at which they are used in school. Thus, it appears that the Rauding Scale produces absolute levels of curriculum material difficulty which are commensurate with what would be expected on a rational basis.

Table 7

Percent of Curriculum Materials, by School Level, at each Rauding Scale Level

	Bormuth Levels				
	1	2	3	4	5
L_d	(N=70)	(N=80)	(N=80)	(N=70)	(N=30)
6	0	0	0	0	21
5	0	1	1	23	70
4	0	7	20	56	6
3	1	18	52	17	3
2	23	56	25	4	0
1	76	18	2	0	0
Total	100	100	100	100	100

Basal Reading Materials

Population and Sample. The population of interest was the basal reading materials used to teach reading in elementary schools. The sample was one of the most frequently used series--Reading 360 published by Ginn and Co. This series contains Ginn Levels 1-13, but Ginn Levels 1 and 2 contain no prose material so they were not included in the sample, i.e., only eleven levels, 3-13, were included.

Rauding Scale. There were three, 100-word passages sampled from each of the 11 basal reading books. The passage samples were chosen by use of the 25th, 50th, and 75th page percentiles, i.e., if a book contained 200 pages, then the three, 100-word samples would be sampled starting on pages 50, 100, and 150, respectively. Each of the 100-word passages were typed on separate pieces of paper and were rated for difficulty by the three raters in Group B in Study 3. A single Rauding Scale value was obtained for each book using the mean of the 9 ratings for each level, i.e., three raters rating three passages each.

Results and Discussion. Table 8 contains the Rauding Scale values for each of the Ginn Readers. Since these books are basal readers used for beginning reading, it is important to note that the Rauding Scale values for Ginn Levels, 3-5, i.e., the first three basal readers, were all Grade 1. If the Rauding Scale values for these beginning readers had been anything but Grade 1, the validity of the Rauding Scale at the lower level would have been questionable.

The Ginn Levels are not officially assigned to any particular grade levels in school, as basal readers used to be. However, many school systems informally adopt certain grade equivalents for the Ginn Levels. The Highland View Elementary School in Montgomery County, Maryland, for example, uses grade equivalents presented in the third column in Table 8. Notice that none of these 11 grade equivalents deviate more than one level from the G_d values. It should not be expected that the Rauding Scale G_d values would perfectly match the Ginn grade equivalents since: (a) only three 100-word samples represented the entire basal reader, (b) there is some error in the Rauding Scale, and (c) there is nothing perfectly valid about the grade equivalents themselves, i.e., as adopted by the school system. Yet, the high correspondence between the Rauding Scale grade difficulties and grade equivalents for the basal readers suggests high validity for the Rauding Scale at the lower levels.

Table 8

Grade Difficulties of the Ginn Readers

Ginn Level	Rauding Scale Grade Difficulty (G_d)	School Grade Equivalent
3	1	1
4	1	1
5	1	1
6	2	1
7	2	2
8	3	2
9	4	3
10	4	3
11	5	4
12	4	5
13	7	6

Newspaper Materials

Population and Sample. Three different sets of newspapers were studied--a random sample of newspapers throughout the United States, a sample of issues from an urban newspaper, i.e., the Washington Post, and a sample of issues from a small town newspaper, i.e., the Leader Journal.

The United States population consisted of one newspaper randomly sampled from each state in the United States during a six month period in 1972. The sampling involved the following procedures:

a. From the U.S. Post Office's 1971-72 National ZIP Code Directory, a single post office ZIP Code was selected from each state using a random number table.

b. Using the telephone directory information service, the name and address of a local newspaper was requested for the town whose ZIP Code was selected above. If the yellow pages indicated that the town had no local newspaper, then the local information operator was requested to give his or her location. Invariably, the operator was located in the same state and in a town close to the target city, so the name and address of a newspaper in the same city as the operator was requested. This procedure tended to favor larger cities and therefore the sampling was not strictly a random sampling of the entire population of city newspapers in a state. However, this sampling procedure probably resulted in a more representative sampling of newspaper readership since the newspapers in larger cities would tend to have a wider circulation than those newspapers in smaller cities.

c. A letter and self-addressed envelope was sent to each of the 50 newspapers, explaining the nature of the research and requesting a recent issue of the paper.

d. Of the 50 requests, 44 responded. For the six non-respondents, a new newspaper was selected to represent the state using the same procedures.

The second newspaper sampled was the Washington Post, the largest newspaper in Washington, D.C. For this population, 25 consecutive daily issues were sampled in the months of January and February of 1972.

The third newspaper sampled was the Leader Journal, a weekly newspaper for a small town (pop. 1,000) located in the midwest--Pierce City, Missouri. For this population, 25 consecutive weekly issues were sampled starting in March, 1970.

For each newspaper, the story located in the top right hand corner of the front page was selected. Then, the first 100 words were counted and the sample ended with the sentence containing the 100th word.

Rauding Scale. Each of the 100 newspaper passages selected were retyped on separate pieces of paper and rated for difficulty on the Rauding Scale by the three Group B raters employed in Study 4.

Results and Discussion. Table 9 contains the percentage of newspapers at each Rauding Scale Level for the three samples. For each of the three samples more than half of the sample was at Level 3, Grades 7-9--United States sample, 52 percent; Washington Post sample, 60 percent; Leader Journal sample, 56 percent. The median grade difficulties for the three samples were 8.2, 9.2, and 7.0, respectively. None of these newspaper passages were at Level 1, Grades 1-3, for any of the three samples. Furthermore, none were at Level 5, Grades 13-15, or Level 6, Grades 16-18.

These data add to the validity of the Rauding Scale. More than half of the passages contained in the front pages of newspapers in the United States seem to be written at Level 3 difficulty, Grades 7-9, while a lesser proportion are at Level 2, Grades 4-6, and Level 4, Grades 10-12. The Washington Post, a sophisticated city newspaper has a median grade difficulty (9.2) which is about one grade higher than the average of all the newspapers in the United States (8.2), and this finding seems to make sense from an intuitive standpoint. Likewise, it

seems to make sense that a small town newspaper in a rural area would have an average difficulty about one grade lower (7.0) than the national average.

Table 9

Percentage of Three Samples of Newspapers at each Rauding Scale Level

L_d	United States (N=50)	Newspaper Sample: Washington Post (N=25)	Newspaper Sample: Leader Journal (N=25)
6	0	0	0
5	0	0	0
4	18	36	8
3	52	60	56
2	30	4	36
1	0	0	0
Total	100	100	100

If the findings had suggested that on the average newspapers were written at Grade 5 or Grade 12 difficulty, this would have been evidence against the validity of the Rauding Scale at the middle levels of difficulty. Or, if the findings had suggested the Washington Post was, on the average, more difficult to read and understand than a small town rural area newspaper, then this would have been evidence against the validity of the Rauding Scale. The Rauding Scale produces difficulty estimates for newspapers which are commensurate with what would be expected from a rational standpoint, and this result adds further to the validity of the Rauding Scale.

Graduate School Materials

Population and Sample. The population of interest is the reading materials used in graduate schools of universities. The materials studied were sampled from the graduate school library of the University of Maryland. This library had the following three reference areas: Humanities, Social Science, and Technology and Science. Ten books were randomly sampled from each of the three reference areas. Then, a 100 word segment was sampled from the center page of each book.

Rauding Scale. Each of the 30, 100-word graduate school passages was typed on separate pieces of paper, and rated for difficulty on the Rauding Scale by the three raters in Group B in Study 3.

The 33 passages in the basal reading study, the 100 passages in the newspaper study, and the 30 graduate school passages in the present study were rated together in the same set of 163 passages. This set of 163 passages was randomly divided into three blocks of 40 passages each and one block of 43 passages. No more than two blocks were rated by any rater on any given day. The order of presentation of the four blocks was randomized for each rater and the passages within each block were also randomized, i.e., shuffled, for each rater.

Since the easy basal reading materials, the newspaper passages, and the difficult graduate school materials were all mixed together, there was no opportunity for raters to develop a response set that would spuriously inflate the validity of the ratings.

Results and Discussion. Table 10 contains the percentage of the graduate school sample of passages at each of the Rauding Scale levels. Half, 50 percent, of the passages are at Level 6, Grades 16-18, 37 percent are at Level 5, Grades 13-15, and 13 percent are at Level 4, Grades 10-12. None of the graduate school passages were at Levels 1-3, i.e., Grades 1-9.

Table 10

Percentage of Graduate School Materials at each Rauding Scale Level

Rauding Scale (L_d)	Percent (N=30)
6	50
5	37
4	13
3	0
2	0
1	0

If most of the graduate school materials had been at the college or high school levels, i.e., Levels 3 or 4, then the validity of the Rauding Scale would have been questionable. However, these data indicate that the Rauding Scale is also valid at the upper levels.

Summary

The Rauding Scale appears to provide absolute values on both the grade, G_d, and level, L_d, scales which are commensurate with what would be expected on the basis of a priori experience with the materials being rated. The results of the basal reader study, the newspaper study, and the graduate school study indicate that the Rauding Scale assigns low level values of difficulty to basal readers as would be expected, and assigns middle level values to newspapers as would be expected, and assigns high level values to materials used in graduate school as would be expected.

The Rauding Scale appears to provide valid grade difficulties throughout the entire school range, i.e., Grades 1-18. This is in contrast to the Dale-Chall formula which does not even purport to discriminate validly below Grade 5 and the Flesch formula which does not provide grade difficulty estimates for the Reading Ease Scores below Grade 5 (see Flesch, 1949).

Study 6

Introduction

Recently, a test has been developed to measure the reading ability of individuals in terms of the most difficult material that an individual can read and understand (Carver, 1974c). The reading test has been anchored to the Rauding Scale so that a Grade 6 ability, for example, means that the individual can be expected to be able to read and understand one-half of the material at Grade 6 difficulty on the Rauding Scale. Thus, a grade ability score on the test means that the probability is about .50 that the individual can read and understand a passage at that grade difficulty. With a reading test that measures individual ability along the same scale as passage difficulty is measured, it was possible to study further the validity of the Rauding Scale.

Method

Subjects and Test. Two forms, A and B, of the National Reading Standards were administered to about 600 students in grades 2-12 of the same school system from which the Teacher Rating$_2$ variable in Study 2 was obtained. The National Reading Standards is the test that provides grade ability (G_a) scores and level ability (L_a) scores that are measured along the corresponding dimension as the Rauding Scale, i.e., G_d and L_d (Carver, 1974).

Passages. The 40 passages used in Study 2 constituted the sample.

Rauding Scale. The ratings of both Group A and Group B from Study 4 were combined to give a composite L_d rating for each of the 40 passages.

Dependent Variable. The dependent variable was the percent of the passages at each L_d which could be read and understood by all those individuals at each L_a. From Study 2, it may be recalled that teachers were asked to decide whether each student in his or her class could read and understand each of the 40 passages. The present study involves a different analysis of these same data. In this study the ratings of the teachers in school grades 7-12 were also used in spite of the fact that they were not likely to be as reliable and valid as the teachers in school grades 2-6. The teachers in grades 2-6 taught the same students all day long while those in grades 7-12 only had them in one or two classes, e.g., English class.

Data Analysis. For each passage and for each ability level, L_a, the percent of the students who were rated by their teachers as being able to read and understand was determined. Then, the mean percent for all the passages at a particular L_a was calculated. This value constituted the percent of individuals at each L_a who could be expected to be able to read and understand the passages at each L_d.

Results and Discussion

Figure 2, which is taken from Carver (1974c), contains the results of this analysis. In general, as difficulty level, L_d, increases, then the percent of individuals who can be expected to read and understand the material decreases, no matter what the ability level, L_a. For example, a Level 3 individual may be expected to be able to read and

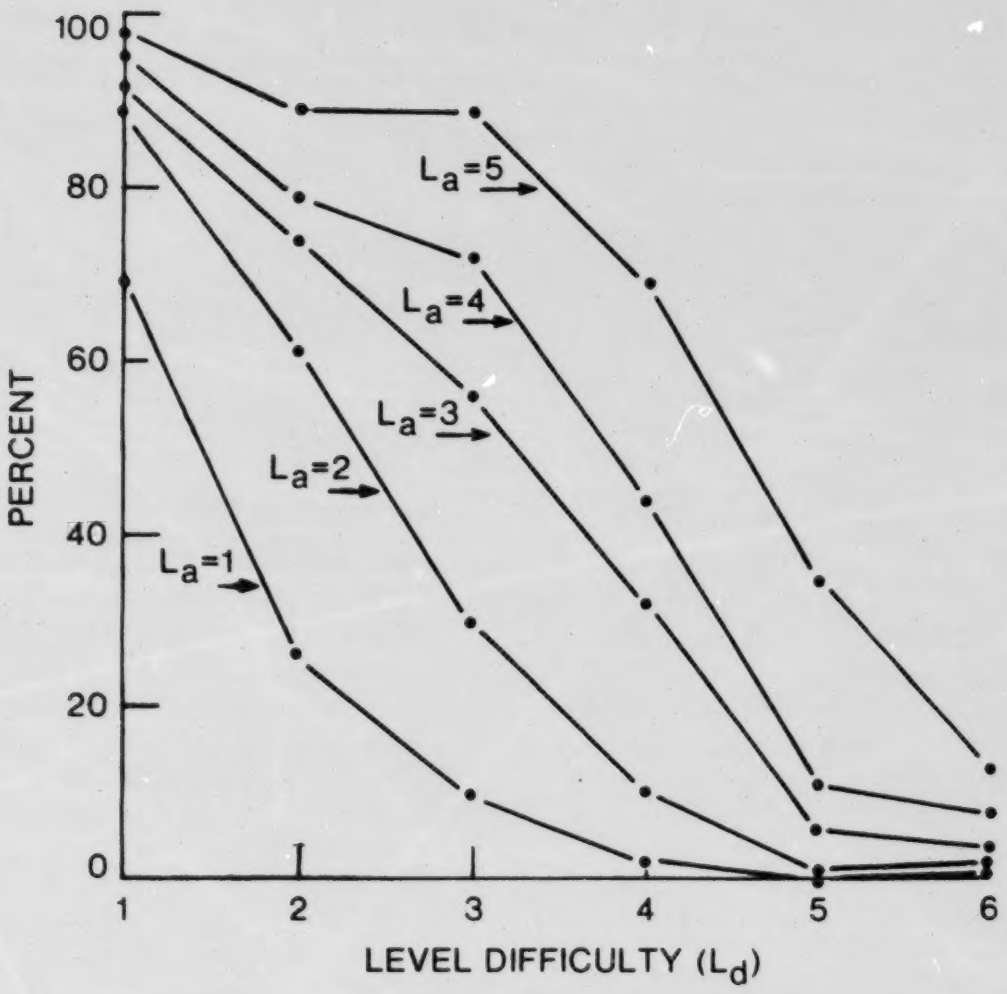


Fig. 2. Percentage of passages at each level difficulty, L_d, which could be read and understood by individuals at each level ability, L_a.

understand about 92 percent of Level 1 materials, 74 percent of Level 2 materials, 56 percent of Level 3 materials, 32 percent of Level 4 materials, 6 percent of Level 5 materials and 4 percent of Level 6 materials.

Suppose a group of Level 1 individuals is given material to read at Level 2. Using the data in Figure 2, it may be estimated that only about 25 percent of the individuals could read and understand the material. Rather than use these point values in Figure 2 to make such estimates, a table has been prepared which involves averages of the data in Figure 2. An average value may be calculated for all individuals who are reading material one level below their own level of ability ($L_a - L_d = 1$), one level above their own level of ability ($L_a - L_d = -1$), etc. These data for the various ability-difficulty differences, $L_a - L_d$, are presented in Table 11.

Table 11

Estimated Percent of Individuals who can Read and Understand Material at Various Ability-Difficulty Differences

$L_a - L_d$	Percent	$G_a - G_d$	Percent
+4	98	+12	98
		+11	96
		+10	94
+3	92	+ 9	92
		+ 8	90
		+ 7	89
+2	87	+ 6	87
		+ 5	83
		+ 4	80
+1	76	+ 3	76
		+ 2	68
		+ 1	61
0	53	0	53
		- 1	43
		- 2	32
-1	22	- 3	22
		- 4	18
		- 5	13
-2	9	- 6	9
		- 7	7
		- 8	4
-3	2	- 9	2
		-10	2
		-11	1
-4	1	-12	1
		-13	1
		-14	1
-5	1	-15	1

If the ability level, L_a, of an individual is two levels higher than the difficulty level of the passage, L_d, then $L_a - L_d = 2$ and the average percent of individuals who could be expected to be able to read and understand the passage in this situation would be 87. These values in Table 11 may be used to estimate the probability that a reasonable degree of understanding will occur whenever the difficulty level of the materials and the ability levels of the individuals are known. So as to make these probabilities applicable to the G_a and G_d scales as well as the L_a and L_d scales, the intermediate values for $G_a - G_d$ have been interpolated and are also presented in Table 11.

It seems reasonable to use information about the difficulty levels of newspapers and the ability levels required to be able to read and understand newspapers to help set standards for minimal functional literacy. In our society, a functioning citizen should be able to read and understand what is happening in our world so as to be able to cast a reasonably intelligent vote and read about job openings, for example. The percentage data in Table 11 may be applied to the United States newspaper data in Study 4 to produce an estimate of the percentage of individuals at a particular ability level who could read and understand newspapers. It has been calculated from these data that only 29 percent of Level 2 individuals may be expected to be able to read and understand most of the material in United States newspapers. For Level 3 individuals, this percent is 54, and for Level 4, it is 75. Thus, in order that the probability be about .50 that individuals will be expected to be able to read and understand newspapers, it is necessary that the individuals reach Level 3, Grades 7-9, of reading ability. This would seem to provide a minimum standard for functional literacy since the probability is low that a Level 2 individual can read and understand newspapers.

Summary

These data further add to the validity of the Rauding Scale by indicating that as Rauding Scale difficulty increases, the percent of individuals who can read and understand the material decreases, no matter what the ability level. More important, however, the data in Study 6 may be used to estimate the probability that about 75 percent of

a passage will be read and understood. This estimate may be made anytime that the reading ability of the individual has been measured along either the G_a or L_a scale and the reading difficulty of the passage has been measured along either the G_d or L_d scale. In the past, ability and difficulty have not been measured in a manner that allowed quantified predictions to be made about what would happen when individuals of certain abilities read passages of certain difficulties.

Conclusions

1. Cloze is a good predictor of the input difficulty of reading material, and as such it is a relatively good predictor of storage difficulty, i.e., the expected degree of difficulty experienced by individual when attempting to read and understand material.

2. The reading-input technique appears to be just as valid as cloze as an indicator of input difficulty, and it seems to have several advantages over cloze as a method for investigating prose difficulty or readability.

3. Average word length, i.e., the RIDE Scale, is valid as an indicator of material difficulty, but it is not consistently as valid an indicator as the Flesch or Dale-Chall formulas.

4. The Flesch and the Dale-Chall techniques were the most valid of the traditional readability formulas studied, but the newly developed Rauding Scale appears to be more valid than these formulas because it is directly sensitive to the idea or concept difficulty of passages.

5. A test has been developed to measure ability along the same scale as difficulty, and it is now possible to estimate the probability of the occurrence of a high degree of understanding when an individual of known ability is presented a passage of known difficulty.

6. The data and procedures introduced in this article may be used to help set minimum standards for functional literacy and to measure progress toward this goal. It appears that Level 3, Grades 7-9, provides a reasonable standard for functional literacy.

PHASE III

The primary purpose of Phase III was to investigate the effect of the interaction between the ability of the individual and the difficulty of the material upon the effectiveness of programmed prose. Programmed prose is simply reading-input material (see Figure 1) that is used to facilitate learning. If the ability of the individual is greater than, less than, or equal to the reading difficulty of the material, what influence will this have upon the facilitative effect of programmed prose?

In order to properly investigate this question, it was necessary to have a good measure of reading ability. As noted in the preceding section, the National Reading Standards measures reading ability in terms of the difficulty of material that an individual can read. It uses a reading-input type of test instead of a reading-storage type of test. Therefore, it seemed desirable to investigate the relationship between ability as measured by the NRS and ability as measured by the reading-storage type of test.

The research in Phase III was divided into two separate studies. Study 1 investigated the relationship between ability as measured by the NRS and ability as measured by the reading-storage test when the difficulty of the material is varied. Study 2 investigated the relationship between the ability-difficulty difference (L_a-L_d) and the effectiveness of programmed prose as a facilitator of learning.

Study 1

Introduction

Individuals of varying ability were administered reading-storage tests on short passages. The passages were at different difficulty levels. The purpose of the study was to determine the relationship between the scores on the reading-storage type of test (RS-test), and scores on the NRS test for various levels of material difficulty. Scores on the RS-test should increase either linearly or monotonically with scores on the NRS test for each level of material difficulty.

Since better than chance scores on the RS-test can be made without ever reading the original passage (Carver, 1973), it also seemed desirable to study the relationship between NRS scores and RS-test scores on passages that were never read.

Method

Subjects. Individuals attending grades 4-12 were administered the reading-storage tests. The school system was the same as in the studies reported earlier.[6] The Ss had been administered both forms of the NRS test earlier. The sample included 398 individuals who were administered the RS-tests and who also had two NRS test scores available.

Passages and Tests. The 16 experimental passages were sampled from the 330, 100-word Bormuth (1969) passages. Four passages were sampled from each of the first four RIDE Levels. The RIDE Levels were used for sampling because the Rauding Scale had not been completed when the study was designed.

Table 12 contains the Bormuth I.D. number for the sixteen passages and the Rauding Scale grades, G_d, and levels, L_a, for each passage at

Table 12
Descriptive Information for the
Sixteen Experimental Passages

Bormuth Identification Number	RIDE Level	G_d	L_d
822	1	3	1
416	1	2	1
015	1	5	2
531	1	4	2
635	2	6	2
723	2	3	1
126	2	10	4
924	2	4	2
327	3	5	2
022	3	8	3
147	3	13	5
133	3	9	3
243	4	12	4
553	4	14	5
053	4	15	5
451	4	14	5

[6] Again, the cooperation of the students, teachers, and administration in the Pierce City, Missouri R-6 School District was appreciated.

each RIDE Level. The Rauding Scale values were determined from Study 5 as reported in the preceding section. Notice that there were 3 passages at Level 1 on the Rauding Scale, 5 at Level 2, 2 at Level 3, 2 at Level 4, and 4 at Level 5.

The RS-Tests on the passages were developed using the standard algorithm for developing this type of test (see Carver, 1974).

Procedures. Each member of each class of students was given a test booklet which contained directions, an example passage, an example test, and six RS-tests. They were told that the testing would only take about 20 minutes and that they could find out what scores they made on the tests. After the directions had been explained, the Ss were given one minute to read the first passage and then four minutes to work on the RS-test on the passage. The next RS-test was administered without any opportunity to read the passage on which the test was based. The Ss were asked to make their best guesses on this test. The preceding procedures for the first two tests were repeated for the following four tests.

The Ss were also instructed to rate their percent of understanding for each passage they read, but these data were not directly relevant to the purposes of this research so they will not be presented in this report.

Design. The first two RS-tests, i.e., one given after reading the passage and one given without getting to read the passage, were exactly the same for all Ss and were regarded by E as practice. The last four RS-tests were all at the same RIDE Level for each S so that every set of four Ss received a different set of four RS-tests. Within each set of four passages at each RIDE Level, the order of presentation of the four tests was varied according to a Latin Square design so that each test was administered once in the four possible order positions. Since there were four possible orders of tests in each of the four RIDE Levels, there were 16 different test booklets altogether. After the test booklets had been assembled, they were stacked in order so that when they were passed out to the Ss, each consecutive set of 16 individuals would receive all 16 possible treatment conditions.

This design provided control over possible practice or fatigue effects associated with the order of presentation.

Data Analysis. Each RS-test was scored, a correction for guessing formula was applied, and finally these scores were converted into a percent correct score. The average of the two NRS Level scores was used to determine each individual's level of ability, L_a. For all the Ss at each L_a, the mean of the RS-test scores on all the passages at each L_d was calculated.

Results

Figure 3 contains the mean percent correct scores on the RS-tests as a function of L_a for each L_d. The values in Figure 3 are for the two RS-tests administered to each S under the reading condition. Notice that the RS-test scores increase almost linearly as a function of the NRS test scores for each level of material difficulty, L_d. The exceptions to linearity are for difficulty Levels 4 and 5 which approach zero at ability Levels 1, 2, and 3.

Figure 4 contains the mean percent correct scores on the RS-tests as a function of L_a for each L_d when the RS-tests were administered under the nonreading condition. Notice that these RS-test scores seem to increase in a positively accelerating manner as a function of the NRS test scores.

Figure 5 contains the gain in RS-test means from nonreading to reading as a function of L_a for each L_d. The values in Figure 5 were calculated by subtracting each value in Figure 4 from its counterpart in Figure 3. Notice that the gain in RS-test scores is generally a monotonic function of the L_a scores except that at the higher L_a scores the gain seems to decrease for the lower L_d values.

Discussion

These data suggest that the L_a scores on the NRS test reflect levels of ability in approximately the same manner as does the RS-test. That is, as L_a increases, then RS-test scores increase no matter what the difficulty level of the material.

These data also suggest that as L_a increases, then the gain in RS-test scores due to reading also increases. The only exception to this generalization seems to involve the situations wherein high ability

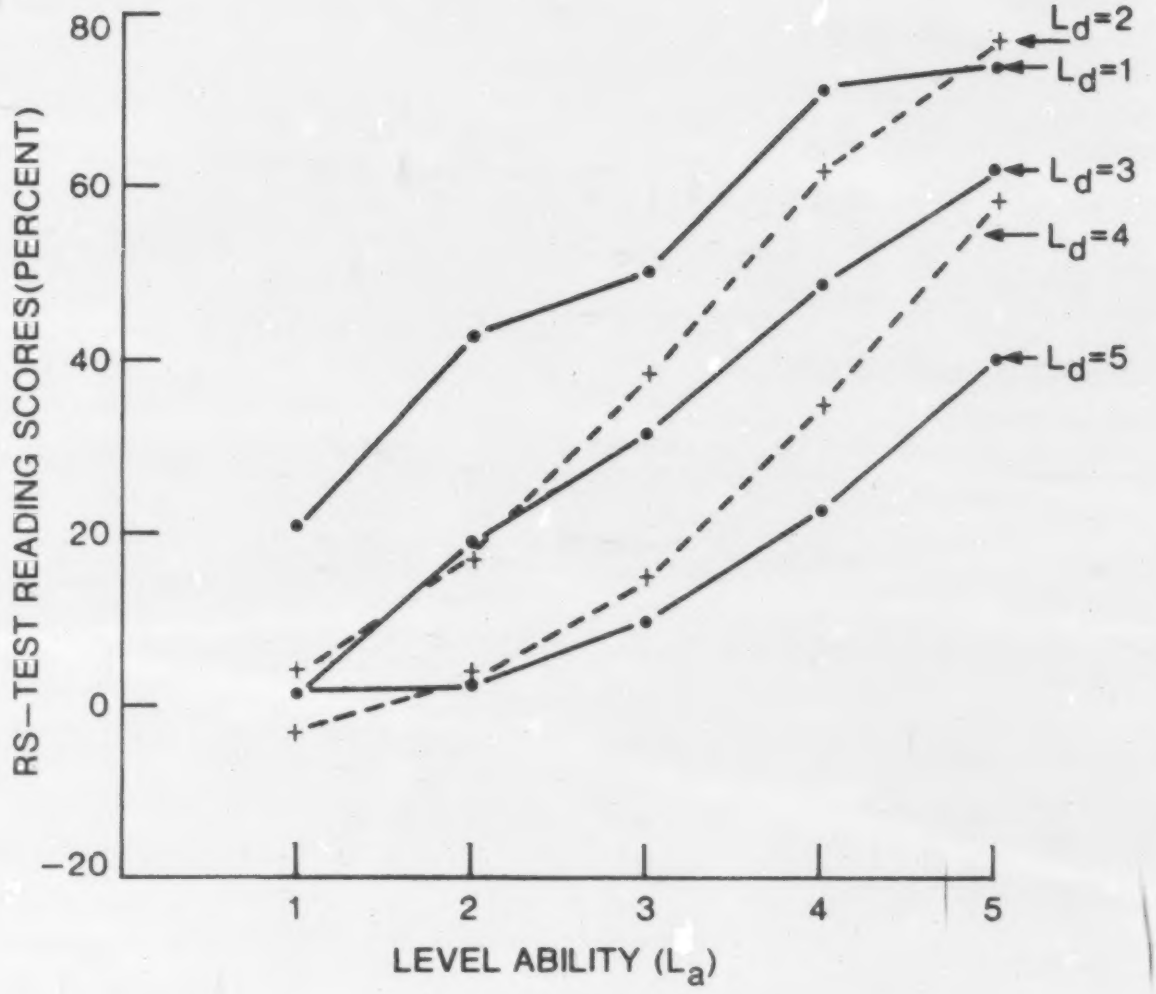


Fig. 3. Mean percent correct on the RS-tests as a function of L_a for each L_d.

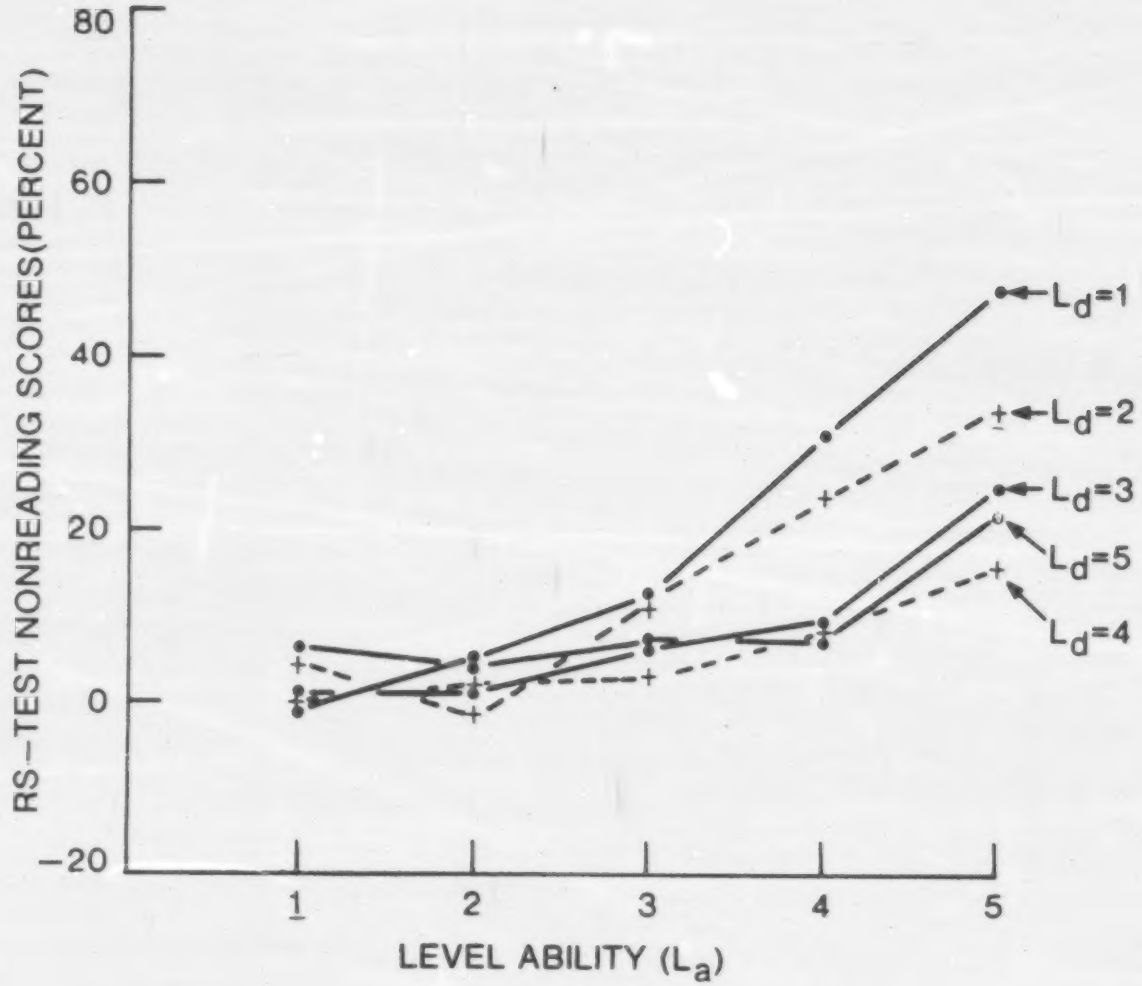


Fig. 4. Mean percent correct on the RS-test, administered under the nonreading condition, as a function of L_a for each L_d.

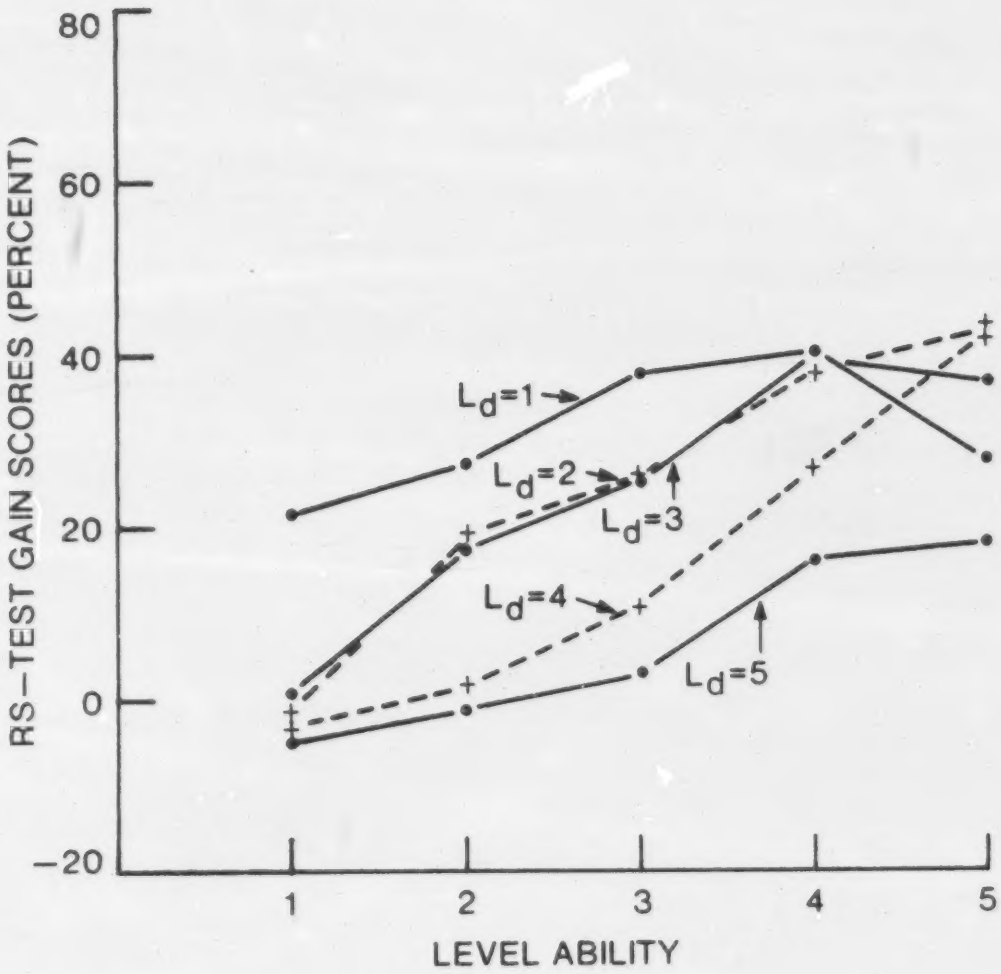


Fig. 5. Gain in RS-test means from nonreading to reading as a function of L_a for each L_d.

individuals, i.e., Levels 4 and 5, are administered RS-tests on low difficulty materials, i.e., Levels 1, 2, and 3. In these situations, a ceiling effect seems to take place wherein the individuals do so well on the RS-tests without ever reading that there is little left for them to gain.

In summary, it appears that ability as measured by the NRS test is highly related to ability as measured by the RS-test. Therefore, there seems to be no reason to develop special RS-tests to measure reading ability when the NRS test provides scores that would likely yield approximately the same results.

Study 2

Introduction

Individuals of varying levels of ability were given a short passage to read and were then tested on what they read. One set of individuals read the short passage in a programmed prose format and the other set read the same short passage in a regular prose format. This study allowed the facilitative effect of programmed prose to be investigated at four different ability-difficulty differences.

Method

Subjects. There were 135 individuals in grades 6, 9, 11 and 12 who took part in Study 2, and they were part of the same group of students who took part in Study 1.

Passage. The 875 word passage represented the first part of a chapter entitled "Sanitation in the Galley and the Wardroom," from the Navy Rate Training Manual, Steward 3 & 2, NAVPERS 10694-D. The passage was about food-borne illnesses and bacterial food poisoning.

The Rauding Scale difficulty of the passage was determined using the same two sets of raters as were used in Study 5 of the preceding section of this report. The passage was divided into 11 consecutive segments of approximately 100 words each, and each rater was given an unlimited amount of time to read and rate the difficulty of each segment.

The mean of the 33 segments for each set of three raters was used to determine the Rauding Scale difficulty of the entire passage. One

set of raters rated the passage at Grade 8 and the other set rated it at Grade 9. Therefore, the passage was estimated to be Level 3 difficulty.

The programmed prose items were developed in the same manner as reported earlier (Carver, 1973) using the algorithm described elsewhere (Carver, 1974d). However, rather than using only upper case letters, as was done in Phase II (Carver, 1973), the programmed prose used a regular format consisting of both upper and lower case letters. There were 175 programmed prose items. The regular prose consisted of the same programmed prose materials except the boxes had already been correctly marked.

Tests. There was a 60 item test on the passage which consisted of 30 RS-test items and 30 paraphrase items (P-test). In a manner similar to the test used in Phase II of this research (Carver, 1973), there were 10, five item, multiple-choice, P-test questions alternating with 10 RS-test questions. One paraphrase question was written on each consecutive sentence. The end of the 60 test items determined the length of the passage.

The Study 1, L_a scores were also used in Study 2.

Time Limits. The time limits for the reading of the passages and for taking the test were chosen so as to provide enough time for almost everyone to finish the programmed prose and to finish the test. Observations during the testing suggested that the 12 minute time limit used for the passage and the 24 minute time limit used for the test were sufficient.

Procedures. The programmed prose booklets were stacked alternately with the regular prose booklets so that when the materials were passed out in each class, approximately half of the class would receive each condition. The programmed prose (PP) group in each class tested was instructed to try to understand what they were reading as they were marking the boxes so that they could answer the test questions. They were also asked to go back and read the material again if they finished early.

The regular prose (RP) group in each class tested was told that the boxes had already been filled in for them and that all they had to do was read the material. They were told that they should try to understand what they were reading so that they could answer the test questions, and that they should keep reading during the entire 12 minutes.

Data Analysis. A correction for guessing formula was applied to both the RS-test and the P-test data, and then the scores were converted into percent correct scores. Mean scores for the PP and RP groups were calculated for each ability level, i.e., each group of L_a scores. The number of Ss in each L_a group were as follows: Level 1, N=7; Level 2, N=11; Level 3, N=37; Level 4, N=55; Level 5, N=25. Since there were only seven Ss in Level 1, the data from this group will not be presented.

Results

Figure 6 contains the results for the 30-item P-test. The PP group did worse than the RP group at ability Level 2, the two groups were approximately equal at ability Level 3, and the PP groups did better than the RP groups at Levels 4 and 5.

Figure 7 contains the results for the 30 item RS-test. Notice that these results are an almost perfect replication of the Figure 6 results except the PP group did slightly better than the RP group at Level 3.

Discussion

The difficulty of the prose was Level 3, and the data may be interpreted in relationship to the ability-difficulty differences. For both tests, the data suggest that programmed prose may be detrimental to reading and understanding when individuals are expected to read and understand material that is one level above their ability level, $L_a - L_d = -1$, i.e., the Frustration Level (Carver, 1974c). Since there were only 11 Ss in the Level 2 group, these data should be interpreted with caution. On the one hand, this result does not appear to be very important because students should rarely be given material at their Frustration Level anyway (see Carver, 1974c). On the other hand, this result is somewhat discouraging from a practical standpoint since it suggests that the technique, as it was used in this research, is not of value when poor readers are reading difficult material, i.e., a situation that often occurs in the real world.

When the ability level equals the difficulty level, i.e., Level 3, there appears to be little or no facilitative or detrimental effect of programmed prose. This result is also slightly discouraging from a practical standpoint. However, since the programmed prose did not

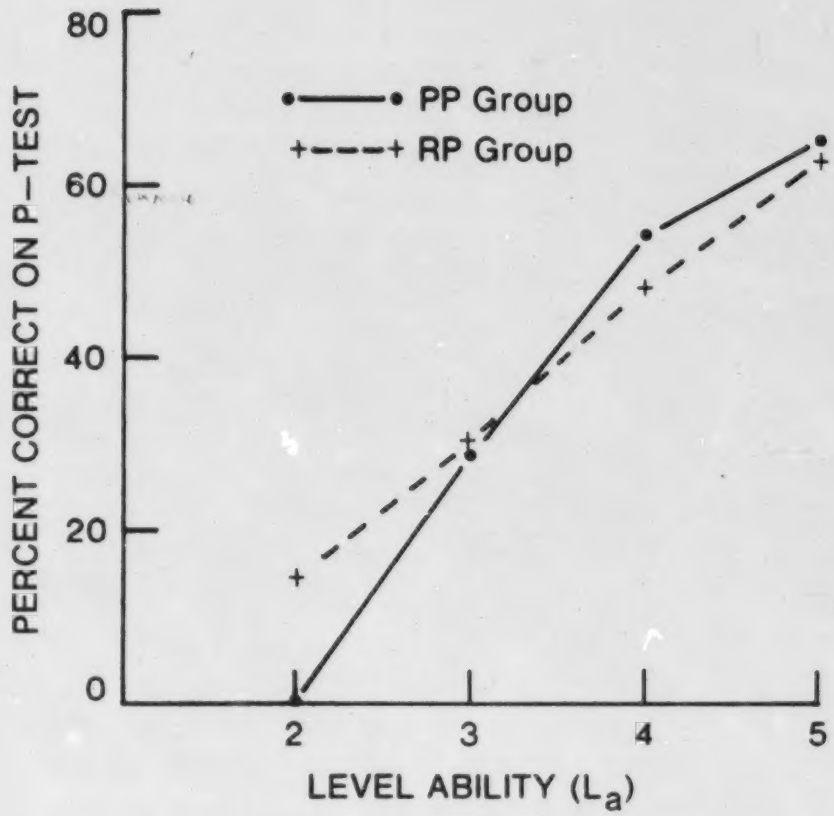


Fig. 6. Mean scores on the Paraphrase Test for each ability level and under both the programmed prose and regular prose conditions.

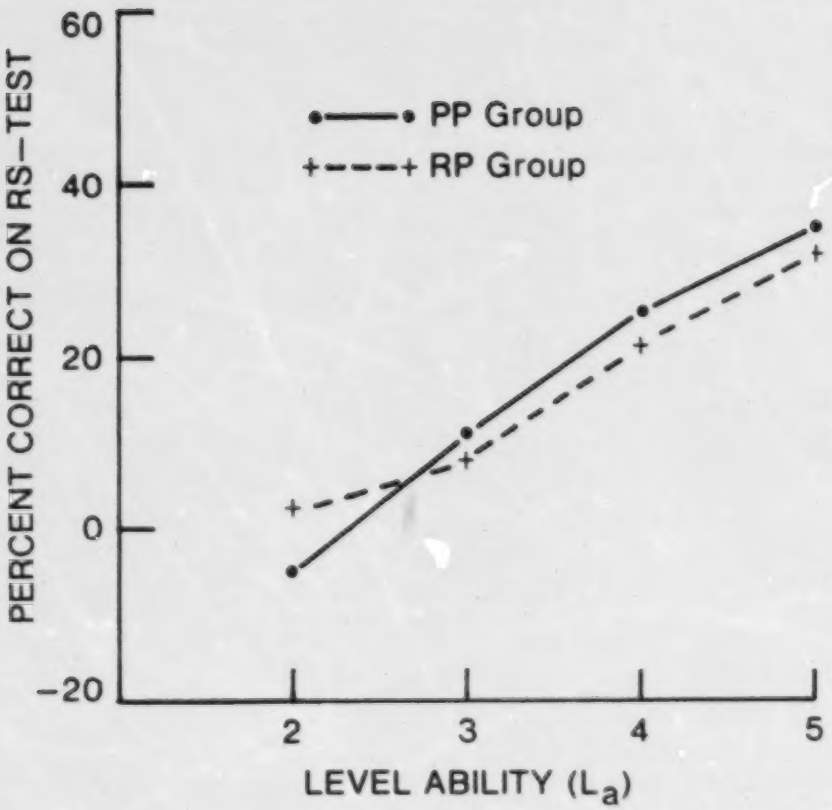


Fig. 7. Mean scores on the RS-test for each ability level and under both the programmed prose and regular prose conditions.

appear to have a detrimental effect on how much was understood while reading, then this would suggest that programmed prose could be used as a reading skill builder with no overall detrimental effect on understanding. It has been suggested that instructional materials in reading should be at one's reading level, i.e., $L_a - L_d = 0$, so these data suggest that programmed prose could be administered as a reading skill builder with no overall detrimental effect upon understanding.

When the difficulty of the material is at Level 3 and the ability of the individual is at Level 4, i.e., $L_a - L_d = 1$, the materials are said to be at the individual's Rauding Level. At the Rauding Level, an individual is expected to be able to read and understand almost all of the material. Thus, at the Rauding Level and above, an individual should experience little or no difficulty with understanding the material; only attention to the task is likely to be a problem. It seems to make a great deal of sense that programmed prose acted as a facilitator at the Rauding Level and above, i.e., Levels 4 and 5.

The motive-incentive conditions in Study 2 were not high, i.e., the students were not getting paid on the basis of how well they did as was the case in Phase II (Carver, 1973c). In Phase II, it was found that programmed prose acted as a facilitator when the motive-incentive conditions were low. Taking these results into account with the present results, it appears that programmed prose acts as a facilitator of attention to the task when motivation wanes and that this effect is beneficial only when attention to the task is the major impediment to learning. When the material is difficult for the individual to read and understand, i.e., the material is at the Frustration Level ($L_a - L_d = -1$), then attention to the task is not the primary problem, and therefore programmed prose probably does not contribute much to prose learning and may in fact be detrimental.

Thus, if programmed prose is to be used as a facilitator of prose learning, it is necessary to have a good understanding of the conditions wherein its facilitative effects overcome its distractive effects. The preceding research suggests that programmed prose facilitates learning when the material is at the Rauding Level, or above, and attention wanes.

Before closing this discussion of the Study 2 results, something needs to be said about the relatively small differences between programmed prose and regular prose in comparison to the test differences attributable to ability level differences. For example, at Level 4, the gain due to the use of programmed prose was only about one-third as large as the gain from Level 3 to Level 4. However, the gain from Level 3 to Level 4 takes approximately 3 years to accomplish, i.e., Grade 8 to Grade 11, so the "1/3 as large" difference might be interpreted as being about as large as what might be expected from one year of schooling. Interpreted in this manner these differences do not seem small. It should also be remembered that these tests were not administered in a nonreading condition to determine how many items could be answered without ever reading. From Study 1 it was found that the higher ability groups can answer more RS-test items correctly without ever reading. Thus, only part of the gain from Level 3 to Level 4, for example, can be attributable to what was gained due to reading. A major part of this gain can be attributed to being able to correctly guess the answers on the test and has nothing to do with what was learned during reading.

PHASE IV

The primary purpose of Phase IV was to investigate further the conditions under which programmed prose might be expected to facilitate learning. The primary facilitative effect of programmed prose seems to come from the fact that it induces an interaction with the prose under conditions wherein attention to the task might be expected to wane. However, there is also an inherent distraction effect associated with programmed prose because the choosing of the correct programmed prose alternative represents a task that is not an inherent part of understanding or storing the information contained in the prose. That is, attention is directed to an aspect of prose learning that is not an inherent aspect of storing or understanding the information. If programmed prose acts as a facilitator only when its inherent distraction effect is outweighed by its attention to the task effect, then something might be done to reduce its inherent distraction effect and thereby increase its facilitative effect. In Phase IV, an effort was made to reduce its distractive effect by providing an item every tenth word instead of every fifth word, as was the case in Phases II and III.

If the above type of analysis of attention during reading is valid for understanding what happens when individuals are induced to learn from prose materials, then it may be that the facilitative effects of another technique, besides programmed prose, could be predicted by a similar analysis of attention (see Anderson, 1970). The traditional study question technique was therefore rationally analyzed and empirically studied as a part of Phase IV.

The study question technique involves the use of questions to facilitate learning from prose. A large group of questions may be presented at the end of a passage, at the beginning of a passage, or a few of the questions may be scattered throughout the passage (e.g., see Carver, 1972). The rationale or theory underlying the use of study questions is usually not clearly delineated. The advantage of using such questions is that it focuses attention upon the prose and requires the individual to interact with the material. The answering of the questions provides feedback to the individual as well as the instructional system regarding how well the individual is doing. Thus, the

attention aspect of the study question technique is similar to programmed prose in that both require an active response on the part of the learner. Thus, both techniques should get the learner involved in the learning process when attention wanes. The study question technique also may act as a distracting influence because the focus may be upon answering the specific questions that are given. That is, an individual may adopt a strategy which focuses upon answering the questions to the detriment of learning anything else in the prose material. Thus, the study question technique may act as a distractor because attention not likely to be focused upon all of the prose. The specific material related to the question should be learned better than normal while the other material not related to the questions is likely to be learned less well than normal (see Carver, 1972).

It may be hypothesized that the study question technique would result in higher retention of the specific study questions than programmed prose and regular prose, but that programmed prose would elicit higher general retention of the prose than either regular prose or the study question technique. This hypothesis was investigated in a series of three experiments.

Experiment 1

Introduction

In this experiment, the regular prose, programmed prose, and the study question techniques were compared when the individuals were reading a lengthy passage at the Rauding Level. To reduce the distracting effect of programmed prose, the material was produced by deleting every 10th word instead of every 5th word. The study question technique involved correspondence course material that is being used in the U.S. Navy. There were a few multiple-choice questions on each page of the material and a special answer sheet provided immediate feedback regarding whether the answer chosen was correct or not. The regular prose was simply the original reading material in an unaltered form.

Method

Subjects. Twenty-six high school males in a large metropolitan area high school[7] responded to an advertisement placed on the student bulletin boards which offered $6.00 for three hours of work. The nature of the work was not explained until the students arrived for the experimental session.

Standardized Test. To control for reading ability, Test 5 of the National Reading Standards (NRS) was administered. This particular NRS test purportedly provides reliable scores for grade abilities 10, 11, and 12, i.e., reliable Level 4 scores.

Prose Materials. The regular prose was a passage that was approximately 4800 words long. It represented most of the entire chapter from which the short passage in Phase III, Study 2 was taken. That is, the short passage used in that earlier study represented about the first one-fifth of the lengthy passage used in this experiment. Some of the subjects covered in this passage on sanitation were the following: chemical food poisoning, personal hygiene, food preservation, and food care. The regular prose material that was given to the Ss consisted of xerox copies of the original eight pages contained in the training manual.

The programmed prose was developed from the regular prose using the same algorithm as was used in Phase III, Study 2. However, rather than using every fifth word as an item, every other item was restored giving one item every tenth word. There was a total of 481 programmed prose items contained in a booklet 30 pages long.

The study question material was developed from the actual correspondence course material used for this training manual. This material consists of a booklet that contains approximately 75 questions on each chapter in the training manual. Each set of these approximately 75 questions has a corresponding immediate response answer sheet. The student is advised to read a chapter and then start answering the questions contained in the booklet. When an answer is chosen, the

---

[7] Altogether, there were three Prince Georges County Maryland High Schools which participated in Experiments 1, 2, and 3. Dr. Victor Rice, Supervisor of Testing and Research, was extremely helpful in coordinating the cooperation of each high school involved. Experiment 1 involved the Friendly High School, and the help given by Mr. Ronald Mortimer, Principal, was appreciated.

student uses his pencil eraser to erase a covering substance to reveal whether or not the alternative answer chosen is correct. If the chosen answer is correct, a "CC" appears. If the chosen answer is not correct, a page number appears and the student should go back to that page and begin reading again until he decides that he has enough information to try the item again. This procedure is to be repeated until each item is answered correctly. The study question material for this experiment was a modification of the original correspondence course material. The questions that were relevant to each page of the regular prose were presented on a page immediately following the prose itself. Since an entire chapter and seventy-five of these questions constituted a task that was too long, given the existing time limit constraints, only the first 50 questions from this chapter were used. The length of the prose passage, i.e., the 4800 words, was determined by terminating the passage after the end of specific material that was relevant to the fiftieth question.

Tests. There were two post tests, i.e., retention tests. One was a 60 item RS-test, similar to that used in Phase III, Study 2. The other test contained the 50 multiple-choice questions (MC-test) that were used in the study question condition.

The 60 item RS-test included 6 sets of items, 10 items per set. The passage was divided into 6 equal segments and a 100-word sub-passage was randomly sampled from each of these segments. The 10 RS-test items were developed from this sampled sub-passage using the standard algorithm (Carver, 1974d).

The 50 item MC-test was developed by simply zeroxing the first 50 items from the chapter in the correspondence course booklet.

Procedure. The entire experiment was described to the Ss in general terms at the outset, and then Test 5, Form A, of the National Reading Standards was administered. The Ss were asked to do their best on this test and were told that their scores on this test would be sent to them with their $6.00 check.

After a 10 minute break, the three types of booklets--regular prose, programmed prose, and study question-- were distributed. The booklets were color coded and stacked so that every set of three consecutive Ss received one of the three types of booklets. Each of the

three groups was instructed to pay close attention to the particular directions that pertained to them, but that they need not pay any attention to the directions given to the other two groups.

The regular prose (RP) group was instructed to read the material carefully and that if they finished the material before the 50 minutes time limit was up that they should go back and read it again. They were told that they would be administered two tests on what they would be reading.

The nature of the programmed prose task was explained to the programmed prose (PP) group at the outset. Then, they were told that they should try to remember what they were reading as they marked the programmed prose items so that they could do their best on the two tests that would be administered at the end of the 50 minute reading period.

The nature of the study-question task was explained to the study question (SQ) group at the outset. They were instructed to read a page and answer the questions on the following page. If they missed any questions they were to go back and study the page again, try another answer, and keep reading and answering the questions until they got all the answers right. Then, they were to go on to the next page and repeat the process until they finished the entire task or the time limit was up.

All groups were told that one of the two tests administered would be a 50 question, multiple-choice test, and that the other test would be different from any test that had ever seen before. All groups were asked to read carefully and were told that their most important job was to do as well as they could on the two tests. All groups were told that their scores on these two tests also would be sent to them.

At the end of 50 minutes, the RS-test was administered using a 25 minute time limit. At this point, the Ss were told that to encourage them to do their best they were going to be paid 4¢ extra for each answer they got correct. An example test was included as part of the directions for this test to assure that each S understood the nature of the task.

After the RS-test had been administered, the MC-test was administered, also using a 25 minute time limit. A bonus was also paid for correct answers on this test. For both tests, the Ss were told that there would be no penalty for guessing so that they should mark one answer to every question.

Results

There were 9, 9, and 8 Ss in the RP, PP, and SQ groups respectively. The number of Ss at Level 4, ability in each of these three groups were 5, 7, and 3. The data from the Level 4 Ss were analyzed.

Figure 8 contains the means for the three groups on both the RS-test and the MC-test. Notice that the PP group scored slightly higher on both tests than did the RP group. The SQ group scored better than the RP and PP groups on the MC-test and worse than these other two groups on the RS-test.

Out of the seven Ss in the PP group, five completed the programmed prose and two did not quite finish. None of the three Ss in the SQ group finished the task.

The mean percent correct for the seven Ss in the PP group on the programmed prose items, after a correction for guessing had been applied was .94 and the scores ranged from .89 to .96.

Discussion

The results suggested that the programmed prose facilitated learning when Ss were reading prose at the Rauding Level under simulated classroom motivation conditions, i.e., Ss were simply asked to do their best. However, since the number of subjects was small and since the size of the facilitative effect did not seem large, a replication study seemed to be in order.

The results also suggested that the study question condition did focus attention upon the questions to the detriment of an overall learning effect. That is, the study question group answered more of the study questions correctly on the posttest but less of the reading-storage questions. It appears that such study questions act to focus attention upon the material directly relevant to the questions while distracting from the other material which did not happen to have a question written upon it.

It was noted during the experiment that everyone in the regular prose group finished reading the material; the average rate of reading was only about 100 words per minute. However, not all of the individuals in the programmed prose group finished the task, and none of the

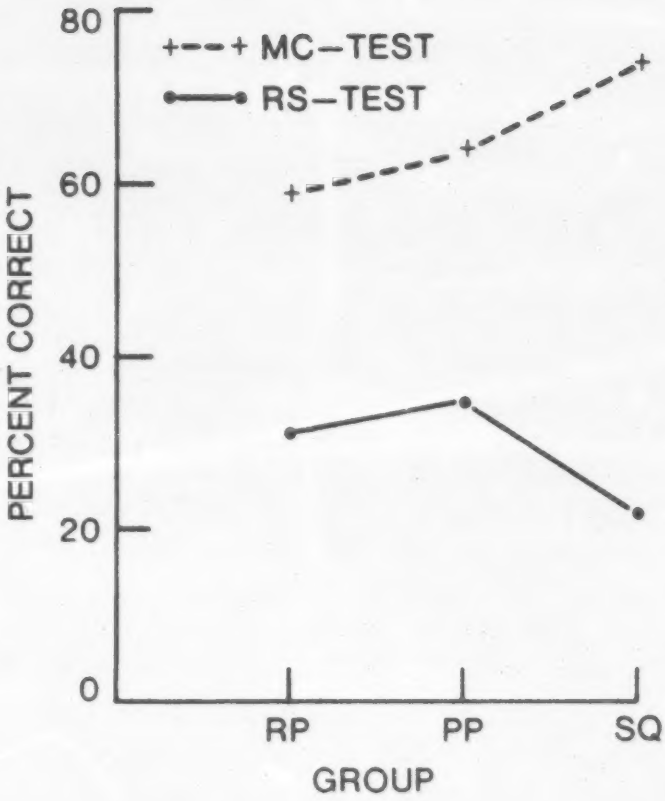


Fig. 8. Mean percent correct for the Level 4 individuals in the regular prose, RP, programmed prose, PP, and study question, SQ, groups for both the RS-test and the MC-test, Experiment 1.

individuals in the study question group finished. Therefore, there appears to be an efficiency aspect of the various methods that should be considered. It appears that the study question condition is the most time consuming of all three techniques while programmed prose requires more time than regular prose.

Experiment 2

Introduction

Experiment 2 was conducted to learn more about the time required to complete each of the three techniques, as well as to determine if the results of Experiment 1 could be replicated.

Method

Subjects. Thirty-six high school males in a large metropolitan high school,[8] different from the school in Experiment 2, responded to the same advertisement as described in Experiment 1.

Materials and Procedures. The same tests and the same experimental materials that were used in Experiment 1 were also used in Experiment 2.

There were a number of procedures that were different in Experiment 2. The PP group was told that they would be given their scores on the programmed prose items, as well as the tests that would take on the material. The Ss were told at the outset that they would take a multiple-choice test on what they read and that they would be administered the test as soon as they finished reading the material. The Ss were not given a bonus for the MC-test, as they were in Experiment 1. The Ss were administered the RS-test last but they were not informed in advance that this last test would also cover the reading material. As in Experiment 1, the Ss were given a 4¢ bonus per each correct answer on the RS-test.

When each person raised his hand signifying that he had finished the material, his booklet was collected and the MC-test was distributed. The E recorded the amount of time on each Ss booklet, to the nearest minute, without letting the Ss know that their reading times were either important or being recorded.

---

[8]The cooperation of Mr. Thomas Moran, Principal of the Oxon Hill High school, was very much appreciated.

The Ss were told that they had a maximum of 55 minutes to work so that they would be administered the test at the end of 55 minutes even if they had not finished.

Data Analysis. Again the NRS test was used to determine which Ss were at Level 4 ability and only the data from these Ss were analyzed. There were 13, 11 and 12 Ss who received the RP, PP, and SQ booklets, respectively, and there were 11, 6 and 9 Level 4 Ss in each of these groups.

Results

Figure 9 contains the results for the experiment. Notice that PP group scored lower than the RP group on both the RS-test and the MC-test. Again, as in Experiment 1, the SQ group scored considerably higher on the MC-test. However, rather than scoring lowest on the RS-test, in this experiment the SQ group scored highest by a very small margin.

The median reading times for the Level 4 Ss in each of RP, PP, and SQ groups were 27.6 minutes, 45.0 minutes and 51.2 minutes respectively. The RP group was reading on the average at a rate of approximately 175 words per minute. All of the 6, Level 4 Ss in the PP group finished during the 55 minute time limit, and their mean proportion correct on the programmed prose items was .95 with the scores ranging from .92 to .98. None of the 9, Level 4 Ss in the SQ group failed to finish the task during the time limit.

Discussion

These data indicate that the regular prose method requires the least time to complete while the programmed prose requires more time than regular prose but less than the study question technique. The study question technique, as it was implemented in this experiment required almost twice as much time as the regular prose.

These data suggest that under these conditions, programmed prose is less effective than regular prose while requiring more time. The study question technique seemed to require about twice as much time as regular prose while being only slightly more effective in facilitating learning, i.e., 38% on the RS-test versus 35% for the regular prose. However, a

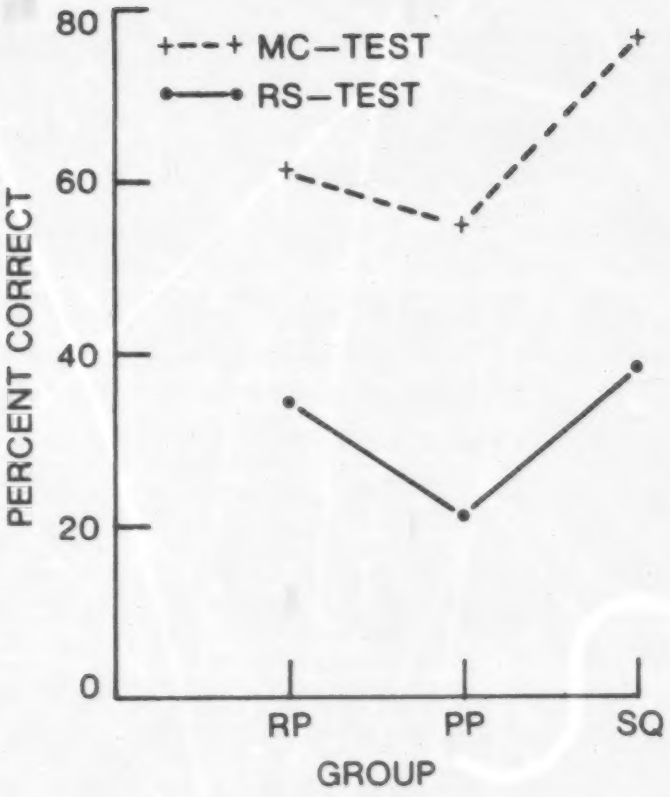


Fig. 9. Mean percent correct for the Level 4 individuals in the regular prose, RP, programmed prose, PP, and study question, SQ, groups for both the RS-test and the MC-test, Experiment 2.

comparison can be made with Experiment 1 that is revealing. The regular prose groups in the two experiments made almost identical scores on the two tests: RS-test, 32% and 34%; MC-test 59% and 61%. These data suggest that the extra time allowed for reading the regular prose was not effectively utilized in Experiment 1. That is, even though the Ss were given extra time to read, i.e., 50 minutes in Experiment 1, they did not use this time effectively. Probably, the Ss read the material once and since there was little incentive to read it again, they did not. Since programmed prose seemed to result in more learning than regular prose in Experiment 1 and less in Experiment 2, this suggests that programmed prose may be effective in inducing readers to spend more time on the reading task. This result suggests that programmed prose is not as effective as regular prose when both are terminated after completion. However, when there is extra time allowed for further reading beyond the first completion trial, the programmed prose treatment is effective in inducing further learning while the regular prose is not.

Experiment 3

Introduction

This experiment was conducted to test the hypothesis developed from Experiment 2 that programmed prose is effective in inducing more study and learning than results from regular prose when sufficient time is allowed for further reading. Experiment 3 was essentially a replication of Experiment 1, but many of the procedures were exactly the same as Experiment 2.

The interpretation of the data in these experiments would be aided if some data was collected indicating what the test scores mean in absolute terms. It is known that some individuals can get some of the answers correct without ever reading, and some individuals may not get some of the answers correct because the items are poorly written. Therefore, a bottom and top control condition would be helpful, as they were provided in Phase I.

Since the data from regular prose groups were relatively stable from Experiment 1 to Experiment 2, these data were not collected again so as to allow the above bottom and top control data to be collected.

Method

Subjects. Thirty-five high school males and females in a large metropolitan high school,[9] a different school from Experiments 1 and 2, responded to the same type of advertisement as described in Experiments 1 and 2.

Materials and Procedures. The same tests and the same experimental materials that were used in Experiments 1 and 2 were also used in Experiment 3.

In general the procedures for the RP and PP groups were the same as Experiment 2 except that the Ss were instructed that if they finished early they should begin reading the material again and keep on reading until the time limit was up, as was the instructions for Experiment 1. Those Ss in the programmed prose groups were instructed to raise their hand when they finished so that their booklets could be collected and a copy of the original reading material could be distributed to them. This procedure meant that upon the second reading, the programmed prose individuals would have the same type of material to read from as the other two groups had. The time limit for Experiment 3 was extended to 60 minutes to give more time to those Ss in the PP and SQ groups to go back and read the material again if they finished early.

The regular prose group was treated quite different in this experiment. During the first 30 minutes of the 60 minute session, this group was administered the MC-test, i.e., a nonreading condition. They were instructed to study the questions and to answer the questions as best they could. They were informed that they would have the next 30 minutes to read the material, and then they would be given exactly the same test again.

Data Analysis. Again, the data from the Level 4 Ss were analyzed. These were 10, 13, and 12, Ss who received regular prose, programmed prose, and study question booklets, respectively, and there were 8, 9 and 5, Level 4 Ss in these three groups.

---

[9]The cooperation of Laurel High School, Mr. Alfred Little, Principal, is gratefully acknowledged.

Results

Figure 10 contains the results for both the MC-test and the RS-test. It may be noted that the mean number of MC-test questions answered without getting to read was 41% and this result for the nonreading condition is signified by the horizontal dotted line in Figure 10. The programmed prose group scored 4% higher than the regular prose group even though the programmed prose group did not have the advantage of studying the test questions before they read. The SQ group answered the most questions correct, 77%, signifying that having feedback regarding which answer is correct and knowing which page to study for each question is a definite advantage with respect to answering these same questions later on a retention test.

On the RS-test, the programmed prose group scored the highest, 4% higher than the RP group, and the study question group scored the lowest, 8% lower than the RP group.

One of the 9, Level 4 Ss in the PP group did not finish the PP task. Two of the 6 Ss in the SQ group did not finish the task.

The mean proportion correct for the Level 4 Ss on the PP items was .97 and the scores ranged from .96 to .99.

In order to provide a more meaningful interpretation of the results, the MC-test scores for all three experiments were converted into an absolute scale. This was a percent gain scale which was derived by letting the 41% nonreading mean score represent 0% gain and the highest MC-test mean score was arbitrarily considered as 100% gain. The highest mean score on the MC-test was 77% for the SQ group in Experiment 3. It seemed reasonable that the Ss who had 60 minutes to study the questions and the text simultaneously, and who received feedback regarding which answers were correct, represent the maximum gain condition. The percent gain scores for all nine means in the three experiments were calculated using the following formula:

$$\text{Percent Gain} = \frac{\text{Mean} - 41}{77 - 41} \times 100. \qquad (1)$$

Figure 11 contains the percent gain scores on the MC-test for the RP, PP, and SQ groups for each of Experiments 1, 2, and 3.

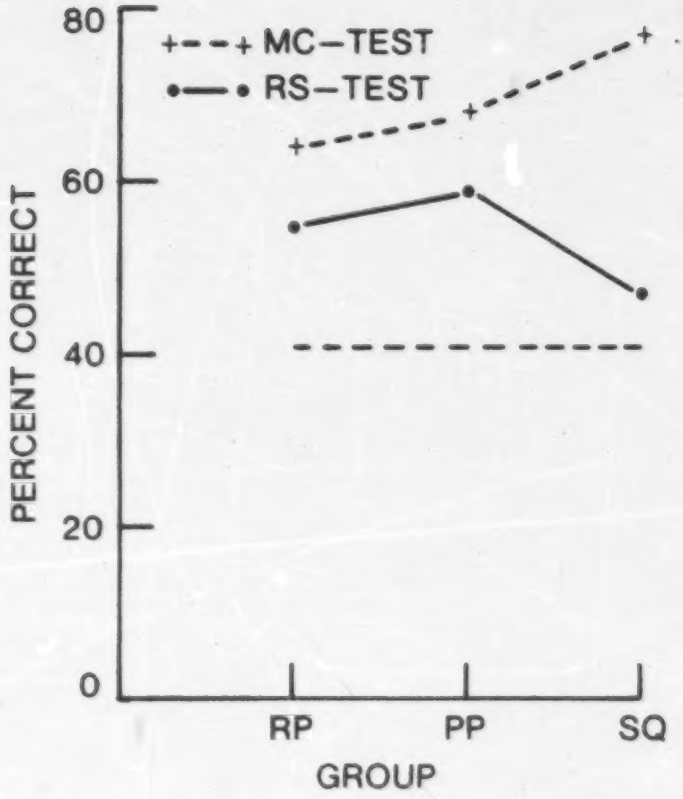


Fig. 10. Mean percent correct for the Level 4 individuals in the regular prose, RP, programed prose, PP, and study question, SQ, groups for both the RS-test and the MC-test, Experiment 3. (Note: The RP group in this experiment received the test questions to study before they read and the mean score for this group is signified by a dotted line.)

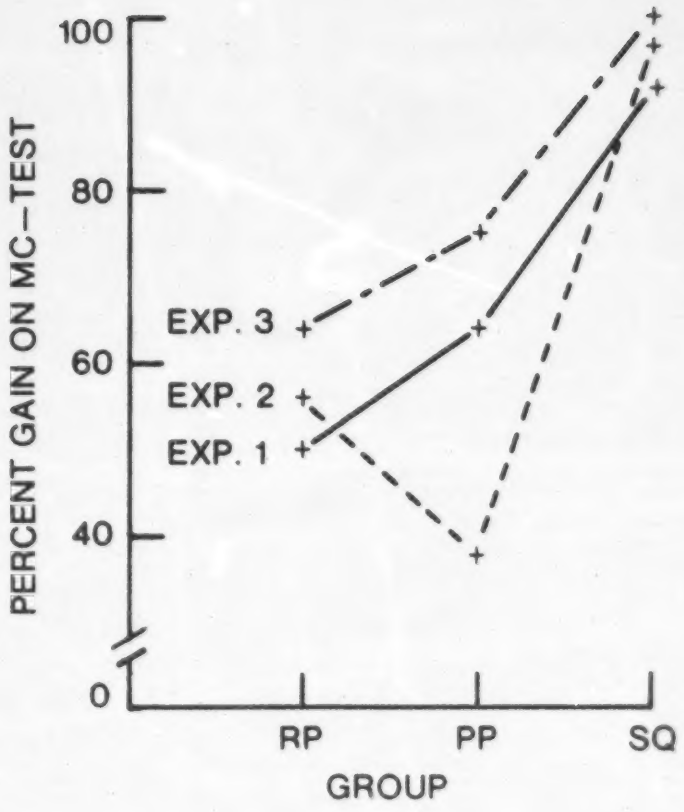


Fig. 11. Percent gain on MC-test for the RP, PP, and SQ groups when the nonreading gain is arbitrarily considered as 0% gain and the highest score is considered as 100% gain.

For the SQ groups, it may be noted that the primary difference between Experiments 1, 2, and 3 was that the time limits were 50, 55, and 60 minutes, respectively. In essence, the 92, 97, and 100 percent gain scores for these three SQ groups represent good replication data for this condition reflecting a slightly higher percent gain for each slightly higher amount of time expended.

For the RP groups, the higher gain score for Experiment 2, 50%, than Experiment 1, 56%, probably reflects the extra 5 minutes given in Experiment 2. In Experiment 3, the Ss had the MC-test questions to read prior to reading the material so it is not surprising that this group had the highest percent gain score of all three RP groups, 64%.

For the PP groups, the lowest gain, 38%, was for Experiment 2 where there was no opportunity to read the material after all of the programmed prose items had been completed. The higher gain score for Experiment 3, 75%, than for Experiment 1, 64%, probably reflects the extra 10 minutes allowed in Experiment 3 for this was the major difference between these two experiments for the PP groups.

If the percent gain for the RP groups in Experiments 1 and 2 are considered as representative of the gain to be expected in this research from regular reading, i.e., 50-56%, then this may be compared to the gain for the RP groups in Experiments 1 and 3, i.e., 64-75%, to get an indication of the effectiveness of programmed prose when it precedes regular reading. If the mean of the two above RP group values, 53%, is compared to the mean of the two above PP groups' values, 69.5%, the amount of increase in percent gain due to the use of programmed prose may be calculated to be 31%. This 31% increase may be compared to the concommitant increase in time found in Experiment 2 from 27.6 minutes to 45.0 minutes, i.e., a 63% increase in time.

Discussion

The data of Experiment 3 tend to replicate the results of Experiment 1. The reading of programmed prose appears to produce higher mean scores than the reading of regular prose when the time allowed for reading exceeds the time necessary to complete the programmed prose. Thus, the facilitative effect of programmed prose seems to be that it induces the reader to spend more time on the task, and thus learn more, when the motive-incentive conditions are not high.

The difference in percent gain between regular prose and programmed prose was around 15-20% on an absolute gain scale. However, this performance advantage of programmed prose was at the expense of a greater amount of time spent on the task. It appears that programmed prose may not be as efficient as regular prose because the 63% increase in time required for the programmed prose was not equalled by a similar percent increase in percent gain, i.e., it was only 31%. Thus, it seems reasonable to conclude that under conditions similar to those in the present research, programmed prose is effective in inducing higher gains in information acquired during reading, but that it is not a highly efficient technique from a time required standpoint.

The attention analysis seemed to provide a way of predicting the outcome of using the study question technique. The study question group always scored best on the test when the test questions were exactly the same questions as the study questions. But, the study question groups scored worse than the two programmed prose groups in Experiments 1 and 3 on the general retention test, i.e., the reading-storage test. This suggested that the focus of attention for the study-question group was upon answering the questions and this distracted the Ss from the general task of reading and understanding all of the material. It also suggests that the study question technique is less efficient than the programmed prose technique for inducing general retention of prose since it required more time to complete and resulted in lower general retention scores.

Discussion

This discussion section will be concerned with the results of all three experiments in Phase IV and how these results may be integrated with the results of Phases II and III.

The general view that emerges from this research on programmed prose is that it acts as a facilitator of learning in certain specifiable circumstances. When individuals are highly motivated and highly competent, regular prose seems to provide the most efficient way of learning. When individuals are not highly motivated, as is often the case in real world learning situations involving prose, then something may be desired which would induce higher levels of learning. When

the level of difficulty of the material is higher than the level of ability of the individual, the learning is extremely difficult. Neither programmed prose nor study questions are likely to help much, at least used in the same manner as they were in the present research. When the difficulty level of the material is below the ability level of the individual and when the time allowed for learning exceeds the time required to complete the programmed prose items, then programmed prose may be expected to induce more learning. When the level of material difficulty equals the level of individual ability, then programmed prose does not seem to facilitate any higher learning, at least as programmed prose was implemented in the present research.

It had been hoped that programmed prose might facilitate learning when the material difficulty level matched the individual ability level, because there are undoubtedly many training situations where this situation is prevalent. Yet, further thought about this situation suggested that it was not reasonable to expect programmed prose to provide higher learning. As noted earlier, it has been suggested that the materials used in reading skill improvement should be at the same level as the individual's ability level. Thus, it can be seen that when the level of material difficulty equals the level of individual ability, we have the optimal situation for improving reading skill, but not necessarily the optional situation for improving the amount learned from prose materials using programmed prose. When the difficulty level of the programmed prose is equal to the ability level of the individual, we might expect the individual to increase his skill in reading this material the first time the programmed prose was completed, but we should not expect the individual to retain much of what was read. This rational analysis suggests that programmed prose might be administered to individuals in this type of situation more than once to facilitate learning. That is, the first one or two trials might be considered as skill trials, i.e., increasing the individuals skill in being able to simply decode this particular material. Later trials with programmed prose may or may not be necessary to induce attention so that the individual can acquire the information or knowledge contained in the material itself.

The theoretical and practical question that emerges from the above discussion is whether repeated programmed prose trials will result in: (a) an increase in the percent of the programmed prose items answered correctly, and (b) an increase in the understanding or retention of the information contained in the prose. It seems reasonable to expect that an individual could become more proficient in reading a particular body of prose material with repeated practice with programmed prose. If a Level 3 individual got 80% of the programmed prose items correct on Level 3 material, for example, it would be anticipated that the individual would be able to increase his percent correct on the prose with repeated trials, especially if feedback were given. This increase would be expected even if each trial contained different programmed prose items. In effect what should happen is that the individual would eventually become a Level 4 ability individual for this particular body of prose. The preceding ideas concerning the interaction between skill improvement and information gained seem to deserve further research because much learning from prose probably does not take place in situations where the ability level of the material exceeds the difficulty level of the prose.

The use of the attention analysis in Phase IV seems to provide a theoretical handle for predicting and explaining the effects of methods used to induce more prose learning. This analysis correctly predicted that the study question technique would result in higher amounts learned when learning was confined to the specific questions used for studying whereas it would produce lower amounts learned when general or overall learning was measured.

From a practical standpoint, the study question technique would seem to be best if one could be sure that the study questions represented the most important information that was to be learned from the prose. Rarely, does one have a situation in the real world where this is the case, however. The study questions usually have not been designed to represent the entire domain of important information. Thus, the danger in using the study-question technique is that important information will not be learned because attention has been focused upon the study questions. Some technique for focusing homogeneous attention would seem to be desirable for most prose.

The programmed prose technique would seem to be preferable to the study question technique because it does focus attention equally upon all the material and it seems to produce higher general retention scores than the study question technique. Furthermore, the programmed prose seems to be a more efficient technique, as it was in this research, i.e., it takes less time to complete and results in higher general retention scores. Yet, programmed prose may not be considered as practical a technique as the study question technique because of the printing costs required to produce the programmed prose version of materials in addition to the regular prose version. This practical disadvantage of programmed prose may be completely eliminated by the presentation of prose training material via a computer terminal, such as PLATO IV (Bitzer, Sherwood & Tenczar, 1973). In fact, with computer capabilities the programmed prose technique would have great advantages over any questioning technique, because the programmed prose could be developed automatically from any regular prose material. The study question technique would require extensive subjective development of questions, an expensive and time consuming proposition. Furthermore, the flexibility of the computer could be used to present programmed prose material only in those situations wherein there was evidence that regular prose was not providing a sufficient level of learning. The practical advantage of using programmed prose to facilitate the amount learned from prose materials seems to lie in the future where computer terminals could be used to present the prose. Basic research will be needed to investigate these ideas about the efficacy of programmed prose as a facilitator of learning when the prose is presented via a computer terminal.

REFERENCES

Aquino, M.R. The validity of the Miller-Coleman readability scale. Reading Research Quarterly, 1969, 4, 352-357.

Bitzer, D.L.; Sherwood, B.A.; and Tenczar, P. Computer-based science education. University of Illinois, Computer-based Education Laboratory, Mary 1973.

Bormuth, J.R. Readability: A new approach. Reading Research Quarterly, 1966, 1, 79-132.

Bormuth, J.R. New developments in readability research. In J.R. Bormuth (Ed.) Readability in 1968. National Council of Teachers of English Research Bulletin, 1968.

Bormuth, J.R. Development o. readability analyses. U.S. Office of Education Final Report, Proj. No. 7-0052, Contract No. OEC-3-7-070052-0326, University of Chicago, March, 1969.

Carroll, J.B. Measurement properties of subjective magnitude estimates of word frequency. Journal of Verbal Learning and Verbal Behavior, 1971, 10, 722-729.

Carver, R. P. Procedures for constructing a variety of information-processing measures appropriate for prose materials. Silver Spring, Md.: Revrac Publications, 1971. (b)

Carver, R.P. A critical review of "mathemagenic" behaviors and the effect of questions upon the retention of prose materials. Journal of Reading Behavior, 1972, 4, 93-119.

Carver, R.P. New techniques for measuring and improving reading comprehension. Washington, D.C.: American Institutes for Research. Technical Report 2/73, February, 1973.

Carver, R.P. Toward a comprehensive theory of reading and prose rauding. Paper presented at the meeting of the American Educational Research Association, Chicago, 1974. (a)

Carver, R.P. Manual for the Rauding Scale Qualification Test. Revrac Publications, in press, 1974. (b)

Carver, R.P. Technical Manual for the National Reading Standards. In preparation for publication, 1974. (c)

Carver, R.P. Revised procedures for developing reading-input materials and reading-storage tests. In preparation for publication, 1974. (d)

Coleman, E.B. Developing a technology of written instruction: Some determiners of the complexity of prose. In E.Z. Rothkopf & P.E. Johnson (Eds.) Verbal learning research and the technology of written instruction. New York: Teachers College Press, 1971. (pp. 155-204).

Dale, E., and Chall, J.S. A formula for predicting readability: Instructions. Educational Research Bulletin, 1948, 27 (Feb. 18), 37-54.

Flesch, R.F. A new readability yardstick. Journal of Applied Psychology, 1948, 32, 221-233.

Flesch, R.F. The art of readable writing. New York: Harper, 1949.

Fry, E. A readability formula that saves time. Journal of Reading, 1968, April, 513-578.

Klare, G.R. The measurement of readability. Ames, Iowa: Iowa State University Press, 1963.

Klare, G.R. The role of word frequency in readability. In J.R. Bormuth (Ed.) Readability in 1968. National Council of Teachers of English, Research Bulletin, 1968.

Klare, G.R., Sinaiko, H.W., and Stolurow, L.M. The cloze procedure: A convenient readability test for training materials and translations. International Review of Applied Psychology, 1972, 21 (2), 77-106.

McLaughlin, H. SMOG Grading--a new readability formula. Journal of Reading, 1969, May, 639-646.

Reed, D.W. A theory of language, speech, and writing. In H. Singer & R.B. Ruddell (Eds.), Theoretical Models and processes of reading. Newark, Del.: International Reading Association, 1970. (pp. 219-238).

Rankin, E.F. Grade level interpretation of cloze readability scores. In F.P. Greene (Ed.), Twentieth yearbook of the National Reading Conference. Milwaukee, Wis.: National Reading Conference, 1971, 30-37.

Smith, E.A. & Kincaid, J.P. Derivation and validation of the automated readability index for use with technical materials. Human Factors, 1970, 12, 457-464.

APPENDIX A--Procedures for Determining RIDE Values

Counting Words. The following guidelines should be followed prior to counting the number of words in the passage.

a. Omit all numbers (e.g., 1973, 5, 1,000,000) and words containing numbers (e.g., 5th).

b. Omit all abbreviations (e.g., etc., ft., a.m., Oct.).

c. Omit each capitalized word, unless it is capitalized because it is at the beginning of a sentence. This includes all proper nouns (e.g., England, Cleveland, Mike).

d. Omit all of the words in titles or names whether they are capitalized or not (e.g., American Federation of Labor, President of the United States, Frederick del Rio).

After the above deletions have been made, the number of words in the passage are counted. In some situations it may be questionable what to regard as a word. In these situations, a word may be described as what is contained between two blank spaces, e.g., "did 'e do it," would be regarded as four words. One exception to this rule involves hyphenated words. If the separate parts of a hyphenated word can stand alone, they are counted separately. For example, mumblety-peg and hold-that-line would be counted as two and three words respectively. Another exception to letting blank space define the beginning and ending of a word involves the dash. For example, "...were looking--it happened..." would be counted as four words, not three.

Counting Letter-Spaces. After the number of words have been counted, the the number of letters (character spaces) within the word should be counted. In general, punctuation marks are not counted, i.e., periods, commas, question marks, etc. However, when these symbols are a necessary part of a word, then they are counted. For example, "don't" has five letters and "re-invented" has eleven letters. A special rule involves words that are hyphenated because they are at the end of a line, i.e., the hyphen is not counted as part of the word since it is an artificial and circumstancial part of the word.

RIDE Scale Value. The total number of letters in the words is divided by the total number of words to get the RIDE value,

$$\text{RIDE value} = \frac{\text{Total number of letters}}{\text{Total number of words}}$$

To prevent connotations of unwarranted accuracy, the resulting value should be rounded to tenths place, e.g., 4.5 and 5.1.

APPENDIX B--Using the Rauding Scale

This appendix will outline the steps for using the Rauding Scale of Prose Difficulty.

Step 1. Administer the Rauding Scale Qualification Test (Carver, 1974) to enough individuals so that there are at least three qualified raters available, i.e., three individuals have passed this test.

Step 2. Each qualified rater should be asked to read a copy of the "Instructions for using the Rauding Scale of Prose Difficulty," contained in Table 13.

Step 3. Each rater should be given the material in question to read and rate on the rauding scale. (Note: There are no standard instructions, procedures, or standard size samples for rating so special procedures and instructions will have to be developed for Step 3.)

Step 4. Calculate the mean of the three ratings for the material.

Step 5. Enter Table 14 with the mean from Step 4 to determine the grade difficulty of the material. (Note: The values in Table 14 represent adjustments in the ratings. These adjustments were made so that the grade difficulties of the Rauding Scale are directly anchored to actual curriculum materials used in specific grades, i.e., as sampled by the Bormuth 330 passages and rated in Study 5. For example, the median Rauding Scale value for all the Bormuth Level 2 passages, Grades 4, 5, and 6, was 6.08 so this value was anchored as Grade 5. The college passages, Level 5, were considered to be Grade 14.5 and the Level 1 passages were considered to represent Grades 2 and 3 so Level 1 was anchored at Grade 2.5. The other anchors were the middle of the Bormuth Levels, i.e., 5, 8, and 11. The other grade values in the table were interpolated using the median for each Bormuth Level.)

Table 13

INSTRUCTIONS
for using the
RAUDING SCALE OF PROSE DIFFICULTY

Only those individuals who have passed the Qualification Test for the Rauding Scale of Prose Difficulty are qualified to use the Rauding Scale.

Remember: Your Grade Level judgments may be anywhere from Grade 1 to Grade 18.

Remember: Your task is to take into account the difficulty of the vocabulary, ideas, and style of a passage, and then choose the Grade Level where you think the average student could read and understand most of the passage.

Remember: You should use the six Rauding Scale passages to determine the grade levels, and you should avoid the temptation to assign grade levels from your own experience.

At the higher Grade Levels, you should assume that the passage is being read by an average student who was majoring in the specialty area covered by the passage. For example, a passage taken from an advanced chemistry textbook might not be understood by an average Grade 16 reader, but it could be understood by an average reader who was majoring in chemistry. In this type of situation, you should assume that the average reader had had the normal amount of prior education appropriate for reading the passage in the specialty area.

Some passages may not apply to the regular school grades. For example, a TV repairman may never go to college so it may seem strange to assign a Grade Level 14 to the kinds of materials that a TV repairman would study from. In these cases, the Grade Levels should be interpreated as reflecting "years of schooling." If, in your opinion, a passage on electronics for a TV repairman is Grade Level 14, then this means that an average reader who studied electronics and TV repair books for two years should be able to read and understand most of the passage. This guideline for technical passages is given to avoid the situation where a passage seems to you to be more difficult to read and understand than Passage 6. Therefore you might consider assigning it a Grade Level 18. You should not do this unless you think it would take an average reader six years of study past high school before being able to read and understand the passage. Even though a passage taken from a textbook on TV repair might use words and concepts that seem to you to be more difficult than those used in Passage 6, you should try to place yourself in the position of a person who has been studying electronics. In which case, you may assign it a Grade Level 15, for example, signifying that you think an average high school graduate could read and understand it after three years of schooling or study.

Next, you will be told the detailed procedures for rating passages and recording your Grade Level judgments. Are there any questions about using the Rauding Scale itself before these procedures are explained.

Table 14

Determining the Grade Difficulty, G_a,
of Material from the Mean of Three Rauding Scale Ratings

Mean of Three Ratings	G_a
1.0- 1.7	1
2.0- 3.3	2
3.7- 4.3	3
4.7- 5.3	4
5.7- 6.3	5
6.7- 7.0	6
7.3- 7.7	7
8.0- 8.3	8
8.7- 9.3	9
9.7-10.3	10
10.7-11.3	11
11.7-12.3	12
12.7-13.3	13
13.7-14.3	14
14.7-15.3	15
15.7-16.3	16
16.7-17.3	17
17.7-18.0	18

RAUDING SCALE OF PROSE DIFFICULTY

Grade 2 — **Trading** — Grade 2

Trading things without the use of money, is called barter.

Did you ever trade toys or cards with friends? If you did, then you, too, were using barter. Such barter has been going on for thousands of years. It is an old, old way of doing business.

At one time, barter was done at big trading centers. People came from miles away to trade things. They brought animals or grain or blankets or straw baskets. Then they traded what they had for something that they needed.

Grade 5 — **Tadpole** — Grade 5

To a young tadpole the world is an amazing place. The waters all about him are filled with tangled forests of bright green weeds all dripping with moss, which wave back and forth with every passing current. From these forests tower great trees, up and up and up until they reach the surface. Some flatten out into huge umbrella-like tops. These are the lily pads. Others go right through and out into the world beyond. These are the cattails and the pickerel weed. In and out of the murky depths of the green forest swim a constant stream of strange unbelievable things.

Grade 8 — **Constitution** — Grade 8

A state may pass reasonable laws to protect the safety and health of its people. But a state law which conflicts with the United States Constitution must give way. The Constitution, for example, safeguards a person's property rights. At the same time, the neighbors have a right to be safe from harm. One right must sometimes be balanced against another.

The Supreme Court has to decide these matters. It holds the delicate balance between freedom and authority...between private property and public welfare...between the states and the nation. When the Supreme Court speaks, it has the final say as to what the law is.

Grade 11 — **Italians** — Grade 11

Gregariousness, curiosity and a fondness for communicating make Italians a universal people, the most universal in Europe. The provincial peasant with whom you share a railroad compartment is completely at home in your company. He may be part of a delegation traveling to a papal audience and it may be his first trip to Rome, but he is nonetheless a citizen of the world. At noontime he will offer you a chunk of his bread with slices of salami and a handful of shiny black olives. He will hand you his Chianti bottle with the crumbs from his lips still clinging to it.

Grade 14 — **America** — Grade 14

America's rise to world power is a consequence of the nation's geographic position, natural resources, and dynamic energy. For the first century and more of national history, however, continental expansion and internal developments largely absorbed the energies of the American people. Every dictate of public interest emphasized the importance of avoiding all entanglements that might involve the young Republic in foreign rivalries and foreign wars. Only with the twentieth century did a rapidly contracting world, impending shifts in the European balance of power, and the growth of American economic and industrial strength create a situation that made impossible a continued aloofness from international affairs.

Grade 17 — **Sternglass** — Grade 17

Yet it would, I suggest, be both unfortunate and inaccurate to depict Sternglass simply as either an irresponsible scientific maverick or as a victim of bureaucratic self-interest. For the question that Sternglass addresses is not a "scientific" one in the rigorous sense of that term. The question "Should the U.S. continue atmospheric testing of nuclear weapons?" is, rather, a question which attempts to mediate between historical circumstance where the available evidence is necessarily incomplete and the need for action on future contingencies. The question dealt with by Sternglass is, in short, an essentially rhetorical proposition. And the disagreement between Sternglass and the governmentally-linked scientific establishment is just another example of the classic "type error"; a dispute arising from alternative interpretation of the same propositition.

AMERICAN INSTITUTES FOR RESEARCH

Corporate Officers

John C. Flanagan, PhD
Chairman of the Board

Paul A. Schwarz, PhD
President

Edwin A. Fleishman, PhD
Senior Vice President

Brent Baxter, PhD
Vice President

Board of Directors

John C. Flanagan, PhD, Chairman

Orville G. Brim, Jr., PhD

Frederick B. Davis, EdD

Paul Horst, PhD

James M. Houston, LLB

Wilbert J. McKeachie, PhD

John D. Montgomery, PhD

Alfred C. Neal, PhD

Richard A. Smith, MD

Research Offices

ASIA/PACIFIC
Bangkok, Thailand
APO San Francisco 96346

KENSINGTON
10605 Concord Street
Kensington, Maryland 20795

PALO ALTO
P.O. Box 1113
Palo Alto, California 94302

PITTSBURGH
710 Chatham Center Office Building
Pittsburgh, Pennsylvania 15219

WASHINGTON
8555 Sixteenth Street
Silver Spring, Maryland 20910